The Role of USAF Research and Development in Information Retrieval and Machine Translation

ROBERT F. SAMSON†

INTRODUCTION

THE United States Air Force has numerous and varied types of data handling problems. This paper reviews some of the developmental approaches and contributions that the Air Force has made toward the solution of semantic-graphic information handling problems. Some of the interesting problems encountered in development of techniques and equipment in this field are presented.

BACKGROUND HISTORY OF ROME AIR DEVELOPMENT CENTER EFFORTS IN THE FIELD OF INFORMATION RETRIEVAL AND MECHANICAL TRANSLATION

In the past four years the Intelligence Laboratory of the Rome Air Development Center (RADC) has learned, through trying experiences, how to get the required movement in this data handling field. Understandably there are many approaches or philosophies, if you will, of how to develop the right synthesis of index, logic, hardware, etc., for any particular informational retrieval solution. The same can be said of mechanical translation (lexicon-logic and hardware). If we allow our minds to review these years and look at the situation as it was when we began our efforts, without today's vast knowledge of hindsight, I believe our approach would be quite similar to the one we took then. We would see the information retrieval problem growing at a staggering rate. The linguistic side was getting some attention, but the hardware, very little. The need and requirements for the Air Force were there and all that remained was to gather funds, select approaches, and secure contracts. I presume you recognize the humor of the previous sentence.

At that time the Air Force started to lend support to various projects already underway as well as to initiate entirely new work in this field. We knew the field needed much development effort, and involuntarily the Air Force took on the role of "catalyst" in information retrieval and mechanical translation developments. Note that I am not saying we were first with most; indeed not—we slipped into a "role" that was important to the Air Force and I believe it has done justice to the problem of both information retrieval and mechanical translation. You will note that I imply the existence of a common problem area in my use of the term "both information retrieval and mechanical translation." Indeed, with the exception of the problem of physically handling documents and their contents, the Air Force Research and Development (R&D) program has been based on the premise that R&D effort in these two areas should be mutually cooperative. To illustrate this part before passing on: it gives a good return for effort expended because the two fields are interrelated, and advance in one usually means advance for the other. For example, if we were interested in information storage and retrieval alone, the Mechanical Translation (MT) field would be suffering for lack of a high-density storage that now seems quite practical. They "complement" one another from a development point of view, not only in hardware as mentioned but also, and perhaps more importantly, from the study of the rudiments of language.

SOME CONTRIBUTIONS BY RADC TO THE LARGE-SCALE INFORMATION RETRIEVAL PROBLEM

Several years ago RADC could not begin to say what type of development catalyst was needed. It could have been in the form of heat generated from "blowing off steam" about the "vast amount of data that must be handled" or it could have been in the form of a stimulating hardware development acting as a catalyst inserted into all the ingredients and by "stirring around to bring about enough agitation to get something done in the field." As mentioned in literature, the old cliché of bewailing the fact that we are being overwhelmed with vast amounts of data and consequently develop only half-vast ideas, was not all correct, although I remember using the expression more than once. We accepted the approach of getting "something" underway and in so doing we became a doer in the field as well as the cause of the needed catalytic actions. From the start we realized we would have to accept the empirical approach; by this I mean a single superior approach was lacking.

We accepted the empirical approach not in total ignorance, for we knew if one was to develop working tools, theoretical analysis alone would be of little help. In our search for new storage media in the field of information retrieval, we came across a high storage density, equipment technique which, when coupled with high read-out rates, could well be the answer to a practical and economical MT look-up or dictionary device. There was one storage medium known at that time that had possibilities of handling densities in the order of 10⁶ bits per square inch; this was the work of King and Ridenour in the use of photographic emulsion on glass disks. The work that followed is now history—the disk photoscopic memory, handling 3×10⁶ bits/square inch, was made feasible, providing us with an extremely valuable empiri-
Fig. 1.

Fig. 2.

Fig. 1.

Fig. 2.

Fig. 1.

Fig. 2.

Fig. 1.

Fig. 2.

Fig. 1.

Fig. 2.

Fig. 1.

Fig. 2.

Fig. 1.

Fig. 2.

Fig. 1.

Fig. 2.
trivial. Although it seems that a device would exist that allowed compact physical storage and efficient retrieval for large-scale document libraries, such was not the case several years ago. Today after some doing instead of speculating, several methods exist. I shall speak of three RADC equipment developments that cover conceptual voids in the field of storage and retrieval. These developments can be broken down into two general categories, both of which depend on environmental and operational conditions when selection is made.

Category 1—The separation of the index from the text. Defined, it is the index of the document removed from the document itself so that the index search is separate from the physical document. The physical document is retrieved by an identification number as a subsequent operation.

Category 2—The combination of the index with the text. The index will also produce the physical document when selection is made.

Under Category 1, at RADC we have the Magnacard Development and the Index Selector. Both these developments come under the heading of technical development, which means in reference to these particular equipments that we are developing "working" tools for experimental use at RADC. In the case of Magnacard, the storage medium is magnetic material deposited on segmented tape, 1X3-inch plastic cards. Engineers at RADC feel that Magnacard has excellent potential for files that require high-speed extraction of information and also where ease of updating and extensive file manipulation by categories is required.

The Document Data Index Set or the Index Search Computer for specialized library as reported in another paper at this conference by Ben Kessel of Computer Control Corporation, is an Index Selector designed for library mechanization. It searches a large volume index and prints out the identification of the documentographic material, etc., that satisfies the search requirements. The Index Selector uses continuous magnetic tape as the storage medium and the scan is serial in fashion.

Under Category 2, that is, index and document text stored together, we find the Minicard program. As one would suspect, this philosophy is based on usage with extremely large files; and, aside from its ability to perform random search, it of course reduces file space and bulk document handling problems considerably. We at RADC consider this as one of the outstanding examples in empirical development approaches, and state without reservation that this technical accomplishment is unsurpassed in the storage and retrieval field. Incidentally, this development has now reached a point where one can say, "It works." It is our sincere hope that large-scale empirical data will be obtained by its application that will give still further impetus to storage and retrieval development. Also at this point, it might be of interest to those in this field that achievements of this kind do not come easily, and I am sure designers and engineers in this field realize fully that 4½ years is certainly a short period of time to develop an aggregate of ten complex equipments having many thousands of interrelated problems involving optics, emulsions, mechanisms, and electronics.

What does all this mean? It means we have mechanized library equipments that will simultaneously give improved operations and serve as tools by which we can experiment with various known library languages and in a relatively short time show the hidden problems in these index schemes themselves. It will also be quite natural to design the index around the logic and structure of the tool. We can prove the worth of indexes by constant evaluation while building a file.

I was asked to include in my paper all the work being done by RADC in the field of information retrieval and MT. In this respect I would like to mention that we are very much involved in the field of character recognition. This interest at first came about through the input problems associated with MT and subsequently considered for all input problems in data handling such as auto indexing, abstracting, etc. We have sponsored a development model which reads one English type font including numerals, both upper and lower case letters, space, and punctuation. We also are under way in developing a Cyrillic character reading machine which will give the MT field a tremendous boost in cutting down the transcription cost.

New York University has recently completed the first phase of a study for RADC on Russian printing matter. This study included such problems as the variety and frequency of Russian type faces and sizes in current use; the reflectance data of the printed type, the reflectance data of the Russian paper, the absorption and reflectance data on inks used in Russian printing, the predominant method of printing, and also the frequency of printing errors.

RADC is also doing other work in the MT field besides developing hardware. A contract with the University of Washington has brought forth a lexicon in the order of 500,000 words with Russian as the "source" language and English as the "target" language. These words will be used on the photo memory of the mechanical translator. RADC scientists are also aiding others in supporting the very interesting work of Dr. Oettinger at Harvard in linguistic work in producing scientific dictionaries automatically. We are also supporting the longer range efforts of the Cambridge Language Research Unit of Cambridge University. This research centers about the use of logical methods utilizing the thesaurus approach in obtaining a translation breakthrough in the multiple meaning problem. Here thesaurus means "an organization of word usage in an ordering dependent on logical content (rather

---

than on alphabetic content as in a dictionary)." These two efforts are supported jointly with the National Science Foundation.

RADC scientists are also aiding in the support of the Research Group of the Center of Studies on Linguistic Activity and Cybernetics, University of Milan, Italy. This work is a continuation of the research studies performed on mental operation and semantic connections. The Research Group is pursuing the approach that man has fundamental order in his thinking process and that these are elements of a correlational net. Taking this correlational structure of thinking and mastering the semantic connections which link the input and output language within this structure, they believe, will be a solution to some of the more difficult problems in mechanical translation.  

We have a development that is completed and although it is classed in the field of information dissemination, we mention it here because it is used in association with storage and retrieval devices. We feel that dissemination exists as an important problem in the continuous flow of data in the field of data handling. This function can be automatized; the equipment referred to is the automatic disseminator jointly developed by RADC engineers and Magnavox Research Laboratory. The disseminator determines what groups are qualified to receive a given document and controls the production and addressing of copies so as to insure that the qualified groups get their copies quickly. The disseminator must determine on the basis of the subject and geographical area of coverage of a given document who is qualified to receive a copy of that document. The disseminator input, as used in one case by RADC, is the flexowriter tape that was used in the Minicard camera for control and code input. The information on the tape is compared to the stored requests in the disseminator as stored in a magnetic drum. The output is tape that contains control data for manufacturing duplicate Minicards based on a match in the disseminator.  


**Some Requirements of the Future**

After this cursory review (and I hope some insight) into information retrieval and mechanical translation development efforts of the RADC, we come to a question of what lies ahead in these two fields. Before I go too far in this direction, I would like to mention that the Air Force has a cardinal interest in the national problem concerning technical information. As can be seen by our efforts, we are going through a "development era" which we feel will have a great influence on the national technical information picture. This is a natural feeling to come from a group that is engaged in developing techniques and hardware such as language research, print readers, automatic language translators, storage and retrieval devices, and disseminators. Equipment such as this will, out of necessity, play an important part in the national picture in both centralized information systems efforts or in decentralized efforts.

Being in the development field, one supposes we should have fine prediction qualities in the semantic-graphic data handling field. Frankly it boils down to studying the trends, following the curves and coming out with the statement that future equipment in these fields should have faster scanning rates, higher excess speeds, greater packing power, lower power requirements, lower cost, etc. However, anyone can make those predictions, but in speaking for a group which has a real invested interest in these fields, we feel the empirical exploitation of developed equipments should be aggressively pursued and that much more should be done in language research for both information retrieval and MT. We feel some effort is "coming about" in this field but many more "bold steps" must be undertaken. Mechanical translation by itself is a language problem, and, by its solution and future use, we only add more literature in the already heavily loaded field of storage and retrieval. Being engineers and scientists we tend perhaps as a group to shy away from the language research side of information retrieval and MT. However, we have slowly learned over the past few years that herein lies the ultimate solution to our immediate common problem.
TWO DISTINGUISH himself from the poor be-nighted man-in-the-street, the computer sophisti-cate is apt to refer to the beasts as "so-called giant brains" or as "lightning-fast idiots." He knows, as do we, the great gulf which separates the human brain from the general-purpose digital computer. Still, the exact dimensions of that gulf are quite unknown, and the desire to show that the hiatus between man and machine is smaller than many suspect impels both the adventure-some and the iconoclastic. The attempt, the successful attempt, to automate one area hitherto con-sidered an exclusively human domain constitutes my topic today.

We are all familiar, if only by hearsay, with the troubles that can beset the best of computer programs if the input to the program is not thoroughly debugged. For a program designed to test the putative behavior of, say, a proposed steam turbine, where the input consists of a scant dozen or so parameters, input debugging is hardly a problem. The situation is quite different for a data-processing operation, particularly when the input is massive, as it usually is. Three alternatives, all unpleasant, present themselves to the supervisor of such a large-scale data-processing operation. He can build a wide variety of error-detecting features into his program, flagging all input errors for subsequent human correction, he can employ a host of human pre-editors to clean up the input, or he can hope that input errors are rare, and let it go at that.

Unhappily, there are many applications where errors are not rare, where the do-nothing solution is obviously frivolous and where, consequently, a sizeable group of humans is necessary, either as pre-editors or as on-line trouble-shooters. Nor is it always the case that the necessary human beings can be clerical types. Certain input debugging calls for sophisticated and knowledge-able practitioners. We are all hopeful—almost all, any-way—that keypunch machines and operators will sooner or later be superseded by character-reading devices and the like. There is no philosophical difficulty in conceiv-ing of typed, printed, or handwritten characters being translated directly into computer language without any human intervention, provided, of course, that those characters were correctly typed, printed, or written to begin with. Suppose, however, that the source char-acters are incorrect. Consider the ingenuity expended in the Post Office just in recognizing all the variations of "Albuquerque." Our Russian colleagues are supposed to be far advanced in the domains of automatic translation and character-reading, but present their machines with a first edition of "Cybernetics," with all its typographical errors, and horrible difficulties would ensue. Our choice, then, is clear. Either we admit that many im-portant data-processing applications are impossible to automate completely, or we find a way to mechanize the human capacity for making educated guesses. We be-lieve that, for some applications at least, we have found a way.

While the techniques we have developed were con-ceived with one particular application in mind, I shall describe them without reference to that application, successful as it was. The principal reason for taking this tack is to be able to present the basic, quite general, features of our method without being tripped up by the special form-fitting required by the actual problem. So, let us be general, and consider any language with which humans attempt to communicate with one another. These may be natural languages, like English or Ger-man, or artificial languages like Esperanto or certain telegraphic codes. There are all sorts of personal rea-sons for communication being difficult—ignorance, dogmatism, poor sentence structure, etc.; however, even if these factors did not exist, all sorts of nonhu-man noise would beset would-be communicators. In-for-mation theory makes much of "redundancy" as an aid in error-detecting and error-correcting when a noisy channel is being used. Indeed, even humans who have never heard of information theory make continual, and skillful, use of redundancy in unscrambling all sorts of garbled communications, whether the trouble be cross-talk in a telephone conversation or missing letters in a crossword puzzle. Without attempting to build a model of the brain, replete with neural nets and such, let us see if we can single out the functions performed by human redundancy-exploiters. If these functions turn out to be performable without recourse to extrasensory percep-tion or to the psychokinetic effect, our automation problem is essentially solved. There remain only the minor problems of collecting all the necessary data, carrying out a rather gruesome programming task and finding a computer fast enough and capacious enough to make our solution practicable. I shall return later to this question of practicability. At the moment, allow me to sketch the functions which, when suitably pro-grammed, allow a general-purpose computer to simulate a redundancy-exploiting, error-detecting, and error-correcting human being.

Rather than jump into a completely general and ab-stract formulation, let me use a concrete illustration. Fig. 1 shows two familiar sights, a correctly prepared mailing envelope and, below it, a somewhat sloppier ver-sion of the same thing. We shall assume at first that a