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 catalystic 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, equipmental technique which, when coupled with high readout 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 10^6 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 3×10^6 bits/square inch, was made feasible, providing us with an extremely valuable empir-

† Rome Air Dev. Center, Griffiss AFB, Rome, N. Y.

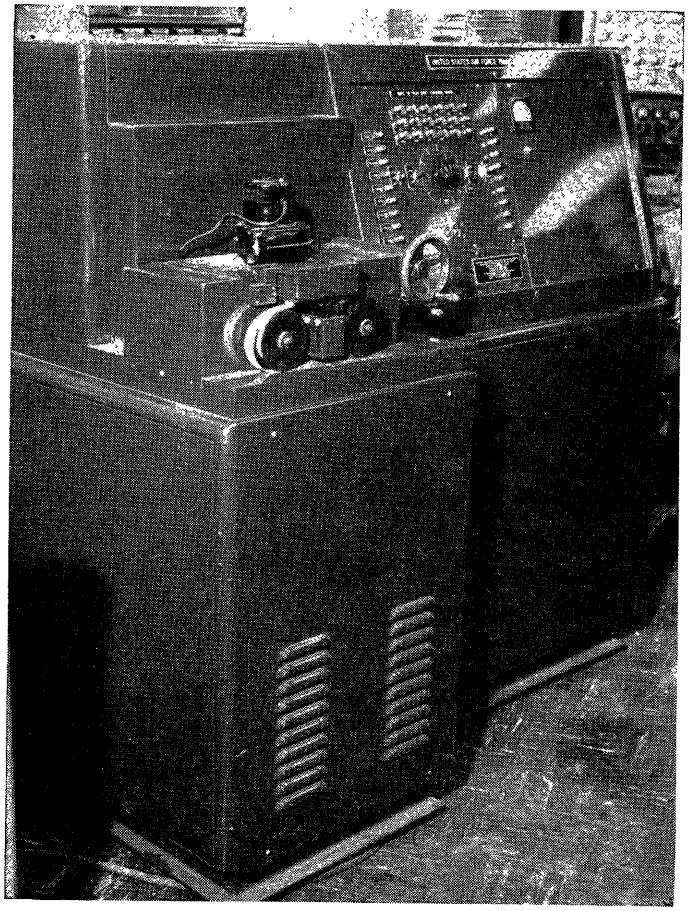

Fig. 1.

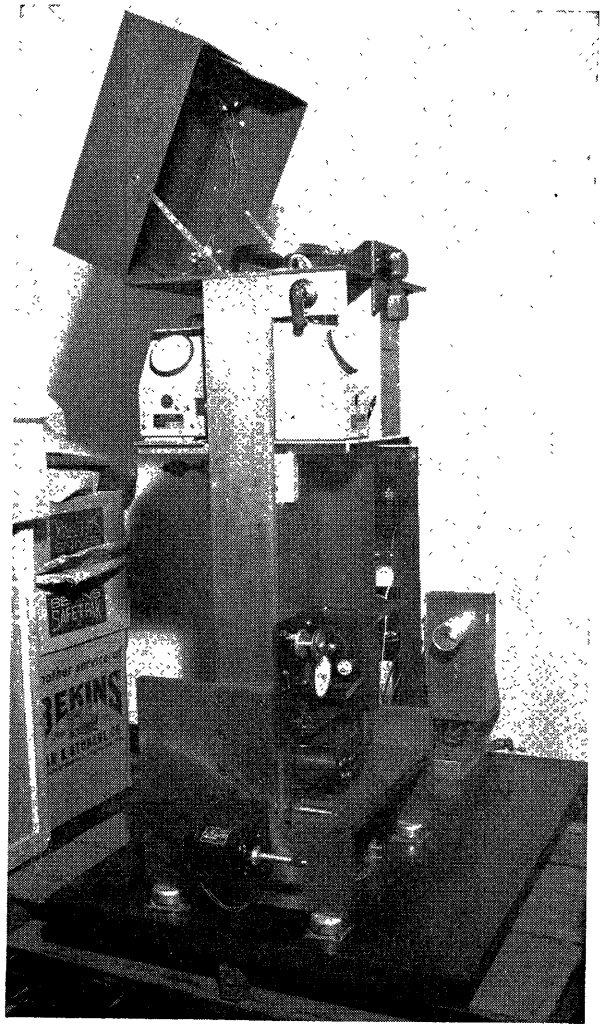

Fig. 2.

ical tool for further research in both information and retrieval *and* MT. This is an example to illustrate the point mentioned earlier, of getting better value in development investment when a development group has interrelated fields. While I am discussing the development of the photoscopic memory, I would also like to illustrate an interesting point. This is of particular interest because it illustrates a sometimes neglected point in developing an equipment that is dependent on a new technique. Referring to the photo disk memory, the equipment necessary to produce a disk was considerable, but of course necessary, if one was to get a high-density storage medium (see Figs. 1 and 2).

These two pieces of equipment by themselves do not represent all the necessary capability required to make a disk, but do show quite clearly the development involved in reaching a goal of practical and economical storage. The point here is that development in these two adjacent fields does not require only development of data handling equipment *per se.* It requires development of all those components that have anything to do with the creating of the media. Actually, the development breakthrough here in terms of what had to be done to produce the required density, was not the disk itself; although this is the end product, it was the precise components that allowed us to make this disk from the raw data on the tapes, thereby providing a facile method of trying many types of stored data. This is not unusual in development programs of this type, but it is the unheralded side. In terms of engineering toil, it represents 70–75 per cent of the work and a substantial portion of the development dollars. Feasibility is a wonderful expression but a tricky term when it comes to development work. It was feasible to reduce a "bit" in terms of laboratory tests—the emulsion always had the resolution—putting the emulsion on optical flat glass had been done—reducing the bits to concentric tracks to disk to show feasibility for a digital store—all this then is "laboratory feasibility," and the cost is quite insignificant when compared to the cost to reduce many millions of bits in a unit of time on a disk at the accuracy required. This had to be done precisely and accurately, and peripheral equipment had to be designed and constructed to make the photoscopic memory workable. The first set has been fabricated and improvements are now underway to perfect the disk-making equipment. The electronic logic used for reading in and out of the photoscopic memory comprise the "other half" of the development.

Now as to the empirical approach in information re-

trieval. 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, 1×3-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 Searcher designed for library mechanization. It searches a large volume index data and prints out the identification of the document-graphic material, etc., that satisfies the search requirements. The Index Searcher 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 $4\frac{1}{2}$ 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[1] means "an organization of word usage in an ordering dependent on logical content (rather

[1] Report on the work of the Cambridge Language Research Unit for the National Science Foundation prepared by Gilbert W. King dated July, 1958.

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.[2]

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.

[2] S. Ceccato, "Mechanical translation," *Automaz. e Automat.*, p. 1, April, 1958.

Some Requirements of the Future

After this cursory review (and I hope some insight) into information retrieval and mechical 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.