[*Eighth Annual Round Table Meeting on Linguistics and Language Studies*, Georgetown University, 1957]

Character Sensing As An Input To Machine Translation

CLYDE C. HEASLY
Intelligent Machines Research Corporation

The need for automatic character sensing extends across the entire data processing field. There are numerous potential applications and, fortunately, a few present commercial applications which are proving the practicality of character sensing and are the source of our most useful experience. The most comprehensive of reading problems are those of reading for the blind and of reading for input to machine translation systems. These problems, in their ideal solution, require machines capable of reading a tremendous variety of material in terms of quality, style and format of the data.

It would seem that ideal solutions of either of these problems are not immediately forthcoming. I would like to describe to you what IMR is doing in some commercial applications and in some laboratory developments. This will give you an idea of what progress one company is making in character sensing. Against this background I will try to indicate what further steps remain in reaching machines of practical utility for machine translation.

In commercial applications, we have reached a goal which seemed far distant a few years ago. We have installed several machines which are now being relied upon for the daily processing of business data. These machines are reading documents of commercial significance and punching what they read into tabulating cards. They process documents at appreciable speeds creating punched cards at rates of from 100 to 180 per minute. A single machine can do the work of a large number of card punch operators; if the equipment fails, work piles up at an alarming rate. But our equipment is working with performance and failure rates that meet commercial operational requirements. One application has been field tested for over a year. Apparently it was not disillusioning, because the installation is being expanded system-wide as the principal means for processing this particular kind of data.

I would like to describe this installation in some detail. It is the application of character sensing to the reading of customer invoices in gasoline and oil credit sales. This equipment is far removed from the problem of machine translation, yet it demonstrates important principles. An important feature of this application is the advance control which was exercised over the nature of the data to be read.

The credit token which is issued to the customer when he is first approved for credit is a plastic Charga-Card* which is embossed with his account number at the oil company's central office. The plastic Charga-Card was developed by Farrington Manufacturing Company who sponsored development of the character sensing equipment. Since the embossed image forms the type from which the account number is printed, any freedom in the design of the embossing can be used to advantage in improving readability of the imprinted invoice. Two steps are taken to improve the resulting image.

First, a maximum reliability type face is embossed into the card. The characters of this type face are formed of line segments which are deliberately made bold to reduce likelihood of broken lines due to light impression. The line segments are arranged to avoid small closed areas which might be filled in by smudge or heavy printing. Each digit of the set of ten digits differs from the other nine by at least two strokes. Since only one stroke-difference between each digit in the set would be sufficient to differentiate between them, the additional stroke-differences amount to a redundancy.

The machine utilizes this redundancy of information to reject digits which are defaced or imperfectly formed. It requires that all strokes which describe the character be present and that all strokes which do not describe the character be absent. If these requirements are not met, the character is rejected.

This use of redundant information is quite useful. The image which must be read is a carbon image which results from a pressure process. Heavy impression and subsequent handling can result in changing the apparent strokes. Light impressions can cause a stroke to be missed. In either case, the machine avoids a mistake by refusing to commit itself.

*Trademark Reg. U. S. Patent Office

At first appraisal, a rejected digit might appear to be as useless as a wrong digit, but this is not the case. The second preventive measure taken in preparation of the credit token makes it possible to restore the rejected digit if only one digit is rejected.

This second step amounts to adding a second kind of redundancy to the account number. An additional digit is included in the number. Thus, if nine digits are required to uniquely identify all customers, a tenth digit can be assigned in whatever manner will accomplish the most good. In this case the tenth digit is assigned according to a simple scheme, the IBM self-checking number scheme. This scheme will tell you the identity of one missing digit if all other digits have been correctly recognized.

The machine reads the account number. If a digit is rejected, the corresponding storage column is allowed to remain empty. The checking circuit responds to each correctly recognized digit. At completion of reading, the checking circuit will lack being satisfied by an amount related to the rejected digit. The checking circuit is interrogated to determine what digit is needed and this digit is inserted into the empty storage column. Thus the number finally punched will be the number which should have been read. As far as correct punching is concerned, a single rejected digit yields the same result as a number correctly read. It would be fortunate if all character sensing applications permitted the luxury of reading self-checking information, but such is not the case.

In other commercial operations, machines are operating successfully even though self-checking data is not available. In such systems the Reader occupies a position quite similar to the human operators it replaces. One such machine is reading the customer's telephone number—which is self-checking—and the amount paid on account—which is not self-checking. In the manual operation which the machine is supplanting, the punched cards are balanced against previously determined batch totals. A similar balancing operation is now used to prove the accuracy of automatic punching by the Reader. If card punch operators were perfect, it would be difficult indeed for an Analyzing Reader* to compete unless the material to be read were perfect. Of course, documents used in the real world are not perfect, but operators are not perfect either. This means that there must be

---

* Trademark Reg. U. S. Patent Office

some means to prove the accuracy of punching by either a card punch operator or a machine reading imperfectly printed documents. It usually works out that the great volume of correctly punched material which the machines can process more than compensates for the procedures necessary to prove the accuracy or isolate errors.

A more advanced machine which comes still closer to the requirements of machine translation input is used to read alphabetic as well as numeric characters. This machine, located at the *Reader's Digest* Condensed Book Club in New York, is used to read a card typed in their Order Department. The field to be read is the customer's name and address, which may have two, three, or four lines. From the first line the name is read and first and last initials selected. The house number is read from the middle line, if there is one, and the city and state are read from the last line.

This Analyzing Reader has difficulty on some material which can still be read by a human. But it does not get bored or tired, and it has "off days" only when components fail. It reads and punches cards at the rate of 150 per minute and this great capacity is particularly useful during the peak volume periods associated with sales promotions. So again, the use of character sensing is proven to be economically attractive, partially because of the ability of the machine and partially because of the intelligent procedures with which it is used.

Perhaps it would be appropriate at this point to discuss some of the means by which Analyzing Readers scan characters and identify what they see. While over five years of development work is behind us, the basic ideas which have evolved can be explained rather quickly.

The first element of the system is the scanner. The scanner can be likened to a television camera in its operation, although its structural details are quite different. A fixed line in space is scanned rapidly and repeatedly as the document to be read passes by. The line of scan, called a "reading station", is parallel to the vertical axis of characters and scanning is so rapid relative to document motion that the scans overlap.

If it is imagined that a capital letter "E" is passing the reading station, the following events will occur: while the vertical line on

the left side is being scanned, a pulse of long time duration will be emitted from the scanner. The duration of this pulse will be determined by the length of the line which caused it. Three pulses will occur during each scan across the three horizontal lines. Finally, after the character has completely passed the reading station, no pulses will result until the next character begins to be seen.

These pulses can be used to identify the character "E" if they are presented to a computer which has been programmed to look for and detect such things as "at least one long pulse", "three properly positioned pulses many times in succession" and "no pulses", provided, of course, that the computer which detects these pulse patterns can also keep track of where they are located, horizontally. This means that the computer can say that the "long vertical line" was on the left side, the "three crossings seen many times in proper vertical position" were in the middle and the "nothing" was on the right.

"End of the character" is used to tell the computer to make a decision on the basis of what it has seen. The things it has seen in this example have been the distinguishing properties of the letter "E", and it will be programmed to emit the signal for "E" every time that unique combination has been seen. In similar fashion, other characters are recognized by the unique combinations of strokes which define them.

The extent to which the details of the character play a part in recognition depends on the requirements of the problem. For example, in most styles of capital "E", the middle horizontal stroke is shorter than the top and bottom strokes, so that at the right end of the character three pulse-scans are followed by a few two pulse-scans. Yet this was not included in recognizing "E" because, in this example, it was not set up as a condition of interest. In similar fashion all other details will be ignored unless routines are set up to detect them. Whether or not the details are implemented will depend on whether or not they are needed in the individual reading problem. In some cases, it may be useful to require that characters be of exactly prescribed size and that certain prominent details peculiar to the type face be seen. In other cases it may be useful to allow considerable latitude as to how long is "long" in a vertical line and to accept three crossings with some variation in spacing. The general method is to instrument whatever characteristics need to be

recognized in order to uniquely identify the characters as found on the actual documents being read.

It becomes evident from the preceding remarks that the machine actually identifies the characters by recognizing the strokes of characters as a code. It is, of course, true that characters are a code and the application of the technique to real problems involves determining how minutely the code must be examined in order to successfully recognize each character in the presence of the noise (smudge, overstamping, light printing, etc.) which accompanies the code.

This reading of code in the presence of noise is related very closely to the question of whether, or perhaps more accurately, *when* IMR can build an Analyzing Reader for MT input purposes. It would appear that many of the documents of interest for automatic translation will carry a great deal of noise. The characters will not be of uniform quality, and quality cannot be improved because the documents are generated by sources over which no control can be exercised.

However, this does not mean that character sensing cannot eventually be applied to MT problems. There are other techniques which can be interposed between scanning and final decoding which greatly enhance the ability to read through noise and through variations in the characters themselves (i.e., subtle changes to the input code). These techniques increase the flexibility with which a machine can analyze a particular stroke and recognize it even though the stroke is imperfect.

Some of these techniques are being applied to a machine now being built which will read the alphabetic capitals, the numeric characters, fourteen punctuation marks and two hand-drawn edit symbols. This machine will read from page copy and punch out what it reads in teletype tape at the rate of 60 characters per second.

The parameters of the problem thus far described are all necessary for machine translation input. The machine falls short of MT requirements in that the material it is to read is prepared under reasonably well controlled conditions on a well maintained printer. It does not read lower case characters, either, but this machine is very much closer to what is needed than the first machine which was described as reading stylized numeric characters only. The simplified

conditions under which the alphameric Reader will perform are perhaps at the halfway point between the simple problems being solved commercially and the more difficult problems of machine translation input.

The use of simplified conditions points to an important facet of IMR's methods. We look at each reading machine in terms of its place in a system. We try to make each element of the system contribute as much as it can to overall performance. Improvements of a minor nature often have profound effects on the output quality of printers and typewriters. Once the other elements have been controlled as well as is practical, we then build the Analyzing Reader of least complexity which will still handle the resulting data.

This approach has led us into problems bearing high economic impetus, where the desire to reduce costs causes minor improvements in document-generating equipment to appear attractive in the light of the end in view. We have at the same time tried to include important innovations in each new machine so that as we proceed from problem to problem our techniques improve and the field of our competence expands.

We believe this to be the method which will cause the fastest progress. Perhaps the steps which we will take on the next few machines will bring character sensing and machine translation closer together. Just as we started reading under restricted conditions and then expanded step by step, so apparently are you beginning your translation efforts in well defined, restricted areas of the source language. As we each expand our scope, the day should arrive when a Reader of less than ideal ability can be combined with a translator of less than ideal ability to solve a practical problem of real and useful significance.