[*Translating and the computer 12: Applying Technology to the Translation Process*. Proceedings of a conference … 8-9 November 1990, ed. Catriona Picken (London: Aslib, 1991)]

Optical Character Recognition

Peter Laurie

My company, Southdata Ltd., is interested in OCR (Optical Character Recognition). I imagine that many of you translate on the computer, and the first problem you have is getting from the paper document in the source language onto a computer so that you can work at it. You could just copy type it, but if you want to save time and trouble you could have a go at Optical Character Recognition. What the OCR tries to do is to run around the page, find bits of ink and decide what they are. Having decided that what it's seen is an e acute or umlaut or whatever it is it then outputs the appropriate computer code. Begin by looking at some of the problems you have with OCR.

This is the process of OCR (a lot of people call it scanning, which is only part of it). You start with a piece of paper and you stick it on the scanner, which is a piece of equipment rather like a photocopier in reverse. It scans the paper but it doesn't output another bit of paper, it outputs a signal to the computer. The signal goes as an image into the computer. Inside the computer, the Optical Character Recognition Software tries to make sense of it. It will put things on the screen to get the operator to help it and then finally it produces texts. The whole process is equivalent to copy typing – it is really nothing more or less than that, just quicker.

What are the problems with OCR? The first and main one is that what the OCR wants is letters that are separate bits of black surrounded by white, which is the definition of a letter. Very few documents are actually printed clearly. They usually look more like a bit of printing which has been photocopied so often it couldn't conceivably be read by Optical Character Resolution Recognition. But if it says, "Can I buy you a drink?" and if

You’re sufficiently close and inspired, you can read it. This is just to remind you how good the eye is: you can read things that machines couldn’t possibly read.

Other nasty problems with typical fonts are “o” and “0”, and capital “I”, “|” and “L”. There is also the problem of “g” and “q”. Now “g” and “q” to the eye look quite different; if you analyse them the only difference is the curve at the bottom of the tail. The rest of the letter is exactly the same. This of course will cause problems because if there is no ink difference in the page the OCR is going to find it difficult to sort out the letter. What we really want is something that not only looks at letters, but looks at the whole word the way the human eye does. You wouldn’t even begin to think of reading a “g” in the word “quality”, but the way the brain does this is unfortunately somewhat mysterious and we don’t know how to imitate it on the computer. We have some ideas but we would need computers perhaps 100 or 1,000 times more powerful than those we have today which can try a word, try a letter, try a word, go to the dictionary, look at the sense of the sentence come back, try another letter. Those are just a few of the problems with OCR.

When you go out and buy an OCR package, and a lot of them are available now, you will find that there are cheap ones and expensive ones. The cheap ones read fast and rather inaccurately, for instance, the scanners you can buy in the Edgware Road in London very often come with OCR packages and they very often produce accuracies of 90% which sounds quite good until you realize that in an average page that means 250 mistakes. So an OCR that delivers anything less than 98 or 99.8% which is three or four mistakes to the page isn't worth having. And the error rate is reflected in the time taken to correct things.

Good OCR means good documents, and if you want to use OCR it is really crucial to start with the best possible source document. You want to get the top copy, you want to shun photocopiers because they mess up the resolution, thus impairing the quality of the print, and if anyone suggest that you should try OCR on a fax, don’t!

Let’s suppose we can read the characters acceptably accurately, and we have a file on the computer of stuff to translate, how are we going to represent them? Now, it’s important to understand how the current computer, particularly the IBM-PC, deals with characters. It has 255 codes, i.e. the contents of one byte, and the byte consists of 8 bits; 32 times 8 is 256 and one of them is blank so that’s 255. So, we have 255 possible shapes we can put on the screen. The current IBM code page 850 is supposed to cater to all conceivable needs, but of course it doesn’t at all. It has the English A-Z in lower case and upper case, it has the numerals and it has some very useful bars and angles to put around little signs on your computer screen. It has a smiling face, and the symbols from a pack of cards, and a few things like capital C cedilla, umlaut and so forth. But a very random selection of possible letters. There is a constant problem in finding computers that will represent the characters you want for your translated document, or in the document you are translating. The problem is actually a bit deeper than this. IBM have simply carried on doing what computers started out doing 30-40 years ago to automate the typewriter. The typewriter has a limited number of characters which appear in a straight line, and the original computer designer thought, well, I'll just give each one a number and that will be that. It will be simple. However, things are not quite so easy.

There may not be the many of you who translate chemical documents, but they are symptomatic of the kind of problem we have, where there is no way of representing a chemical formula, because in the set of 255 possible letters all printed out in a straight line, you can't have some of them smaller and higher and others smaller and lower. So what has to happen is that the printer has to be given some kind of exotic code which makes it change the font, jump up half a line, print, and change font and jump down. This is a real stumbling block, and an immediate example of where the computer screen and keyboard representation don't match the real world. But things get worse... Take a really horrific equation, something which mathematicians write absolutely naturally and other mathematicians read naturally. Printers can handle it, we don't ask how, a lot of coffee and cold towels I would imagine, but the computer really doesn't have a mechanism for dealing with this. We are a long way away from the idea of letters appearing in a neat line across the page. It's a kind of two dimensional maze, and the OCR will recognize those characters, but what it can do with them is a complete mystery. I'd love somebody to tell me how to represent them. Complex equations do point up the fact that contemporary computers are not built to deal with real life.

Next, a telex cannot be represented at all! There is no way of dealing with small caps... In fact, I think there is a big gap to be jumped over before computers are really as capable as the printed word and the written word.

This has been a rather depressing talk, I'm afraid, I hoped to leave you felling inspired but I'm afraid all I've pointed out is trouble. But these are problems that must be solved before we're going to get much further.

AUTHOR

Peter Laurie, Managing Director, Southdata, Voysey House, Barley Mow Passage, London W4 4PT.