Information Traces in the Computer

Natural language, whether spoken or written, is not exact. In daily life it does not need to be. Everyone understands the request: 'Would you please move a bit to the side' – distance and direction do not need to be specified but are evident from the situation. Anybody who says: 'Please transpose yourself through 65 cm in a north-easterly direction' is either aiming for humorous effect or demonstrating that he is dreadfully pedantic, if not mad.

Only for very special applications is the imprecision of natural language unacceptable. Mathematics and logic form one example; in this field natural language has been progressively displaced by symbolic notations and symbolic languages in which each symbol has an unambiguous meaning.
A much newer area in which the imprecision of natural language gives problems is that of information retrieval with the aid of computers. This involves very large files – containing titles and summaries of publications, for example – stored in the computer's memory. The user wishes to extract from the file those publications that he needs for this work. The incompatibility between the imprecision of the language and the computer's highly exact mode of working then comes to the fore. Almost all systems require the material for which a file is to be searched to be identified by means of keywords. The keywords must be specified exactly, since most systems will not tolerate alternative spellings or typing errors. If the keyword is POLLUTION and the user enters POLUTION, most systems will not treat this as meaning the same.
In many cases a single keyword is not enough, more are needed to specify the query in sufficient detail. The user then has to know how he must interrogate the file and what combination of keywords he must use to give himself the best chance of obtaining a good answer. This applies with even more force if not one file is to be searched but several, for there is often an enormous variation in the way the information is formatted.

Terminals enable the user to give more details about his query, by means of a dialogue with the computer, if he doesn't like the answers he is getting – but they do not remove the problem of operation. Even with a terminal the assistance of an expert with a thorough knowledge of the systems is generally indispensable. Before letting the query loose on the computer file the user, together with the expert, must work out the best formulation.
But the ideal would be for the untrained user to be able to sit at the terminal and express his query in whatever way he thinks is most appropriate, obtaining an answer like: 'Publication X seems most relevant to your query; is that what you are looking for?' The computer would then be adapted to the imprecision of human language, instead of requiring humans to conform to its exactness.

The Central Organization for Applied Scientific Research in the Netherlands ('TNO') has recently completed a program that goes a long way towards fulfilling this wish. It has been christened FUZZIE and can be used in whatever way is desired: both keywords and queries in a natural language are accepted. This new program is based on two ideas. The first may be regarded as the fundamental principle of all information retrieval systems: that there will be a certain degree of correspondence between a query and good answers to it.
Somebody looking for publications on 'Glanders in horses' may confidently expect the words 'glanders' and 'horse' to occur in the titles and summaries of good answers. All systems currently in use work on this principle, but the user first has to select keywords that he expects to be present in the relevant publications and enter them into the computer. The designers of FUZZIE were of the opinion that this step should be performed by the computer too. This led directly to the question of how the computer can determine what correspondence there is between the query and a good answer.
As far as we know the basic units used by all existing information retrieval systems are whole words. It has always been tacitly, and incorrectly, assumed that the machine must learn, as it were, to 'understand' our language.
The machine has been regarded as operating on 'concepts', and in language these are carried by words. But a machine cannot 'understand', it can only determine that a certain combination of letters is stored in its memory. Words are essential for input and output, but not inside the computer. This line of thought led to the use in the TNO program of basic units that are smaller than words; each sentence is automatically split up into groups of three letters, called '3-strings'. The word FUZZIE, for example, is converted by the computer into basic units FUZ, UZZ, ZZI, ZIE. Complete texts are split up in the same way. For storing texts in the computer's memory a tree structure is used. Each separate publication proceeds, as it were, along one particular branch of the tree structure, leaving a trace of 3-strings behind it. Each query, too, is converted in this way into an information trace of 3-strings.

For information retrieval the computer compares the information trace of the query with the information traces of the publications that are stored in the memory, selecting those information traces that most closely correspond to the information trace of the query. The degree of correspondence – the 'importance value' of the answer – is calculated by the computer and presented to the user, so that he can see immediately whether query and answer correspond reasonably well or not.
In reality the program's structure is more complicated than has been suggested here. It functions with the aid of an inverted file, and the calculation of the importance value takes account not only of syntactic correspondence but also of semantic correspondence – that is, the content.
FUZZIE has several advantages: first and foremost the simplicity of operation. The user enters his queries in natural language and need not know anything of the internal structure of the system. What he does need to know is which language – Dutch, English, etc. – is in use in the system. Inappropriate query formulation is immediately evident from the answer and from the size of the importance value.
New users find it easy to play with the system and in this way quickly master the necessary routine. This puts information retrieval within the grasp of all sorts of institutions, such as medium-sized and small businesses, for whom it was until now too expensive.

The system offers the facility of holding complete texts in the computer's memory. The capacity of the current version is very large, being sufficient to accommodate nearly ten thousand million documents. Partly in consequence of this, new areas of application are being opened up for information retrieval. In some fields a user has little or no use for titles and summaries, but always requires the complete text. An obvious example is jurisprudence, in which the exact wording of statutes, precedents, and judgements is of prime importance. The quantity of jurisprudence is increasing rapidly – partly because of EEC legislation – and automated information retrieval has been needed for a long time. This has not proved feasible with current systems, but FUZZIE is expected to change all that. A Dutch publisher of legal reference books is investigating the possibilities in collaboration with TNO.
Another possible application is automatic translation. Texts in one language and their translations in others can be stored in memory by the trace method. With FUZZIE, the computer can then convert a sentence in one of the stored languages into each of the others. Such automatic translations should not be expected to be perfect, they will always need correction by people.

Perhaps FUZZIE can help solve the problem of the polyglottism of the information explosion. TNO is certain that other applications will be found as this new program for information retrieval becomes more widely known.

Literature

T. de Heer, Quasi comprehension of natural language by means of information traces, *Information Processing and Management*, in the press.
T. de Heer, Een nogal slordige zoekstrategie, *Open*, 10, Sept. 1978.

Further information from:
Institute TNO for Mathematics, Information Processing and Statistics, T. de Heer, P.O.Box 297, 2501 BD The Hague, Tel. +31 70 824161