In Defense of Natural Language

Vincent E. Guiliano, Arthur D. Little, Inc.

Natural language has evolved to the present point in a way optimally to meet the needs for two-way human communication. This holds both in terms of the general folk argot and the specialized versions of English, German, Japanese, etc. used in scientific, technical and other specialized discourse. Many properties of natural language, such as the frequency distribution of word usage, the use of syntactic nesting, etc. have been much studied in the last twenty years. They seem to serve certain valuable functions in providing a matched communications interface between people, given human's information processing, memory and communicating capabilities. For example, natural language provides adequate redundancy, error detection and correction capabilities, as well as for making best use of limited temporary memory capacity and for enabling communication within limits between individuals with quite different personal memory data bases and cognitive frames of reference.

I argue in this presentation that English is an optimal query language for certain kinds of interactions with certain types of data bases, specifically when the data base consists of textual rather than highly organized numeric data, when the performance desired is retrieval of relevant passages of text or documents. My chain of argument runs roughly as follows:

(1) English has evolved to be in some sense optimal for communication interchange among humans, including query-answer processes of various kinds; it is the language with which humans have had the best and longest experience, both phylogenically and individually, are most comfortable and function the fullest and best.

(2) There is increasing evidence that the main bandwidth and information processing limitations in using advanced interactive human-machine information storage and retrieval systems lie often on the human side rather than the machine side. This consideration suggests use of natural language for querying data bases if at all possible.

(3) Elementary techniques for interactive storage and retrieval of natural language texts, for example, coordinate retrieval of text words, have been used for over a decade and work well enough to provide the bases for large scale natural-language text retrieval systems now in commercial operation. The query languages usable with such data bases usually consist of Boolean formula expressions involving logical connectives, English words and parenthetic delimiters. Advanced techniques allowing the use of free text queries, such as use of computed "associative" thesauruses to facilitate recall independently of initial query vocabulary selection, were extensively researched in the 1960's, found to work well but have not yet been tried on a large commercial scale. Such systems are based on queries consisting of simple interrogative sentences: "Give me information about ...," machine display of vocabulary alternatives; they rely on good searching results emerging not from using complex queries, but rather from a simple interactive process. I am not considering systems here that require handling highly structured queries, such as "Is it true that ...?"

(4) The applicability of natural language as a data base query language for the purposes mentioned has been massively misunderstood for over a decade due to several confounding factors, including:

(a) Confusion between such text retrieval objectives and "fact retrieval." Many small scale experiments were conducted in the 1960's with "query-answer systems" for handling highly structured queries with very limited but also highly structured data bases. Such experiments at first seemed to be very promising but when looked at closer were not easily extendable.

(b) The use of information storage and retrieval as a cover for basic research on Linguistics. A very vocal group of researchers insisted that the one and only approach to use of natural language for any retrieval application is the use of transformational grammars, in spite of -- or perhaps because of -- the fact that the problems of such grammars have only grown more and more complex and esoteric over the last fifteen years.

(c) The absence until recently of available remote-access computer environments backed up with adequate mass memories to enable economical implementation on natural language text retrieval systems on other than very small scale experimental bases.

I predict: (1) the next decade will see a variety of natural language data base information storage and retrieval systems come into day-to-day practical operation, and (2) the tendency in the query language structures used with such systems will be more and more towards use of natural language queries.