[From: *Information retrieval and machine translation*, ed. Allen Kent. International conference…September 6-12, 1959, Cleveland, Ohio; part 2 (New York: Interscience Publishers, 1961)]

CHAPTER 40

In Defense of English

VICTOR H. YNGVE*

Massachusetts Institute of Technology, Cambridge, Massachusetts

"The essential purpose of literature searching is to locate those documents within a collection which have a bearing on a reasonable question. A reasonable question must be defined as any serious question—of obvious or potential significance—posed by persons who have... important reasons for desiring the answer to the question."

"Every statement of a technical article should be retrievable in any frame of reference established by the search request; e.g., a trash incinerator having means for collecting and removing unburned residue may be a reference for the recovery of precious metal from goldsmith's apparel by burning the apparel and recovering the ash. In other words, one must be able to retrieve any facet or aspect of every statement."

The preceding paragraphs are excellent expressions of our goal. The first was written by Perry and Kent in their book "Tools for Machine Literature Searching" (p. 33); the second, by Andrews and Newman in a report "Activities and Objectives of the Office of Research and Development in the U. S. Patent Office" (1).

To me, the most significant word in both of these paragraphs is the word 'any'. We want to locate documents bearing on any serious question. We want to retrieve <u>any</u> facet or aspect of every statement in <u>any</u> frame of reference.

In the Patent Office, the examiner's object in making a search is to determine whether anything in the file constitutes a valid anticipation of the claim that he is examining. If there are things in the file that are relevant, he wants to be able to examine them so that he can make decisions about patentability. If he cannot find anything in the file that is relevant, he wants good assurance that there is indeed nothing there. The examiner is interested in obtaining an answer, yes or no, for every one of the over 2,800,000 patents, to the question: Does this document contain anything that might possibly be relevant to the application that I have at hand?

It is physically impossible for the examiner to look at every one of the patents. However, in certain cases he may not have to look at

*The work presented here was supported in part by the National Bureau of Standards.

every document. If he finds a valid reference, he may not have to look any further, even though there may be other valid references in the file. Because the file is very large, the patent classification system has been set up to help him find valid references. The system is of great use because, although the examiner must still look at the patents individually, his search is first directed to groups of patents in which a reference is likely to be found if there is one. In many cases his directed search through a manageable number of patents is rewarded by the discovery of one or more valid references.

But if the examiner is not able to locate a valid reference in any of the obvious places in the classification system, he is faced with a problem. How can he know for sure that there is no valid reference anywhere in the file? After he has availed himself of all the help that the classification system can give him, he has either to give up the search and assume that there actually is no valid reference in the file, or he has to start searching blindly through the whole file—a very unrewarding, if not actually impossible, task.

The present classification system is a great service when valid references are accessible in the file, but it is quite inadequate when the search is directed to something new that cannot be located by using the classification system. In this case, it is impossible to make an exhaustive search because it is humanly impossible to search the entire file to find the answer to a novel search request.

Many of the systems that are now used for searching a large file can be represented by the diagram that is shown in Fig. 1. The documents of the file are first abstracted according to a particular scheme of abstracting. The abstracts are then encoded by an essentially mechanical but frequently not mechanized process. The questions must be interpreted in the light of the kind of questions the system can answer. The interpreted (or abstracted) questions are then en-

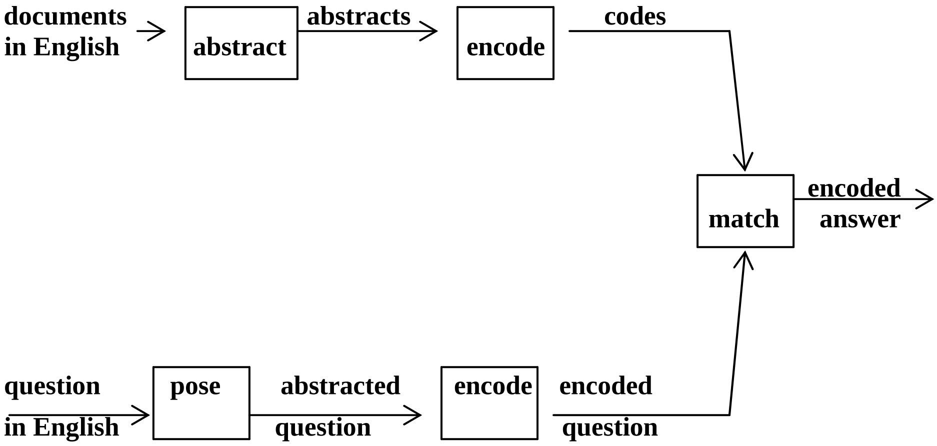

Fig. 1. Idealized search system.

coded, and the resulting code is matched, manually or mechanically, to the abstract codes. For our present purposes we can ignore other steps in using the system, such as the steps of interpreting the encoded answer; feedback to the user, who may ask other questions; and feedback to the abstracter, who may reabstract or reclassify a portion of the file.

Although some of the systems set up on the scheme shown in Fig. 1 have achieved a considerable measure of success, many of their most serious difficulties can be traced to weaknesses in one or another of the steps. The step that we have called abstracting amounts to a careful manual searching of each document for answers to certain predetermined questions. Besides the element of human fallibility there are two serious problems associated with abstracting. The first concerns the sheer bulk of the file and the manpower and time required to encode it. This amounts to a very large investment, and one that is not lightly made. The other difficulty is, however, even more serious. It is due to the necessarily limited scope of the questions answered during abstracting. The system can give direct answers only to those questions that the abstracter has considered and effectively answered. The system fails when the user comes with a question that has not been foreseen and that consequently has not been searched for and encoded ahead of time. Some people have thought that a great intellectual effort could produce a classification system capable of answering any question, no matter from what angle it had been asked. In the opinion of the author this is probably impossible. One cannot search for the answer to a question that has not in effect been answered in the abstracting process. Abstracting, by the very nature of the concept, reduces the bulk of the file to be searched and consequently leaves out much that is revealed in the documents themselves. The material left out can never be retrieved by the system. The difficulty in principle with the concept of abstracting, involving as it does an inevitable loss in information and a concealing of the answers to many possible questions, leads to the concept that it may be necessary to search the file directly by machine. In the past it has been considered unrealistic to consider searching English (or French or German, etc.) text directly. There have been a number of good reasons for this. But now there is some hope that such an objective might be realizable.

In the old days, the most promising machines for retrieval purposes were punched-card machines, or photoelectric machines having many of the characteristics of punched-card machines. The only operations that machines could do then was look for exact matches and perform elementary switching operations. But today, machines have a great deal more flexibility. The electronic digital computer is capable of carrying out complicated processing of data. We no longer have to live with the exact match limitation.

Figure 2 shows a postulated search system for directly searching English text for answers to questions expressed in English. It can be seen that the search program requires as input, besides the English

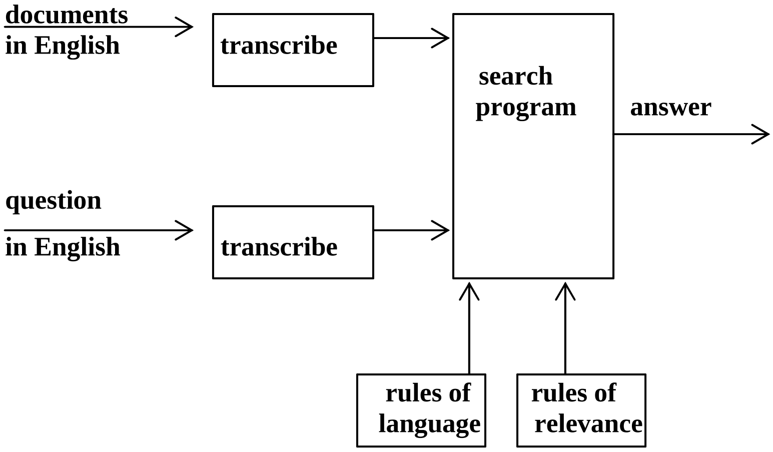

Fig. 2. Postulated search system.

text and the question in English, two subsidiary files, one of them containing the rules of English and the other containing the rules for judging the relevance of text as answers to questions. We shall not discuss here the possibility that the storing and retrieving of answers to questions might be more economical than researching the file. Neither shall we discuss the economics that are involved in searching first those parts of the file most likely to contain answers. We are, here, interested only in the possibility of searching English text directly.

The full explicit nature of the rules of language and the rules of relevance are, of course, as yet unknown, but we already know something of their nature. The rules of language are yielding to linguistic research. The work that is being done on hierarchical classifications is an example of our growing knowledge about rules of relevance. Much remains to be done.

The advantages of a system that can search text in English are great enough to warrant the expenditure of a considerable effort to develop such a system. First of all, there would be no abstracting, classifying, or encoding to be done manually. For large files, this would be a great saving. Second, there would be no loss of information between the text and the file to be searched. This is an important result of the eliminate tion of abstracting. It, in turn, makes possible the third advantage--that it is easy to update the system. No reclassification or re-examination of text would ever be necessary. If, for example, it is found that a certain chemical has insecticidal powers, it is not necessary to re-encode all references to this chemical to allow retrieval. It is merely necessary to alter the rules of relevance, part of the “grammar” of the system. This advantage can be obtained even in a system that does not search text directly, if the rules of relevance are stored in a separate file instead of in the document codes (2, 3). The fourth advantage of searching English text directly is that there is no essential limitation on the scope of allowed questions imposed by the system of encoding. If the information desired is expressed in the text, it is, in principle, possible to find it if the rules of the language and the rules of relevance are sufficiently well-known.

Let us look at English, and other languages, and ask whether or not they are well-suited to the task. We have a set of compatible languages that have never been equalled as a medium for the expression of ideas. They are both highly standardized and widely known. Experts in their use are everywhere. These languages are beautiful instruments of precision. The existence of partial synonyms makes possible precise shades of meaning in a compact notation without extensive qualification. With extensive qualification, one can be as precise as necessary. But at the same time the languages do not pedantically insist on precision where it is not needed. The degree of precision is left to the good sense of the user. The extensive use of constructions where the meaning is more than the sum of the individual meanings of the separate words makes for a great economy and conciseness of expression. We are not limited to a number of meanings equal to the number of words in our vocabulary. Much meaning is carried in the context, making possible the multiple functioning of words without confusion. The resulting homonyms result in a smaller vocabulary and a consequent economy. It is possible to express essentially the same thought in a number of ways, depending on the situation. We are not held in a strait jacket that would allow us only one way to express each thought. There are many special short expressions available for frequently expressed meanings.

These many advantage are bought at the price of complexity, yet the complexity is not so great that children cannot easily learn the languages. It should be possible, by an appropriate research effort, to deduce the rules of language in explicit form so that machines can be explicitly instructed on how to use the languages. A good start has already been made.

One way of proceeding is by a step-by-step approach (2) that will take us from where we are in our understanding to where we want to be. This approach operates within the traditional concept of abstracting, encoding, posing, and matching. Each step, however, makes the codes closer and closer to English and the encoded question closer and closer to the original question. Each step will be a “dialect” of English that we understand thoroughly and can search with something more than a simple match. Each dialect will incorporate within it more of the features of English—as many as we feel we understand thoroughly at the time. Eventually the dialects will become essentially English itself and we will no longer have to work within the traditional framework. We will then be searching the file directly for answers to questions in English.

REFERENCES

1. Andrews, D.D., and Newman, S. M., “Journal of the Patent Office Society,” Vol. 40, No. 2, Feb. 1958, pp. 79-95.

2. Yngve, Victor H., The Feasibility of Machine Searching of English Texts, “Preprints of Papers for the International Conference on Scientific Information,” 1958, Area 5, pp. 161-169.
3. Leibowitz, J., Frome, J., and Andrews, D. D., Variable Scope Patent Searching by an Inverted File Technique, “Patent Office Research and Development Reports” No. 14, Nov. 17, 1958.