[*Proceedings of the National Symposium on Machine Translation*, UCLA February 1960]

Session 8: INFORMATION PROCESSING AND LINGUISTIC ANALYSIS

FROM TEXT TO TOPIC IN MECHANIZED SEARCHING SYSTEMS

Thyllis Williams
Itek Corporation

We now conceive of many information processing tasks which we would like to accomplish in computer-based systems. Mechanized translation and mechanized searching are two of them. I suppose that all of us here at the National Symposium on Machine Translation are hopeful that we will achieve some measure of respectable success in the case of both translation and searching and that we will know, scientifically, how we managed to do it.

In speaking about translation and searching, I shall limit the scope of my remarks to the translation of scientific and technical documents and to searching with respect to collections of such documents. Given this limitation, one can make some rather definite statements about similarities and differences between the two tasks.

For both translation and searching, the primary initial inputs are documents consisting at least partly of discursive text, presented somewhat faultily and inconsistently in a writing system which mirrors only darkly the segmental and suprasegmental regularities observable in spoken language. For both, there is in one dimension and on one level a common practical objective, which is to aid scientists, technologists, and others associated with them, in their exploitation of the cumulative record; not just to avoid unwanted duplication but also to bolster, enrich, and in subtle ways modify their present and future work. For both, also, there can be a legitimate focus on one dimension of "meaning", the cognitive or informative dimension.

The fact that translation and searching are differently concerned with informative meaning, in relation to text, is what distinguishes them most markedly. In translation, one wants to go from text to text (i. e. , from text in source language to text in target language) by procedures which transfer informative meaning completely and reliably. In searching, one wants to discover connections between text (i. e. , source text) and topic. It is true that we consider the topic to be rendered or renderable in a text or utterance; but the status of the topic is such that its rendition is in some sense secondary

or derivative. In translation, a primary question is how a given complex of informative meaning can be communicated accurately and conveniently in language-system . Here, any way that can be found will suffice. In searching, a corresponding question is how, in what ways, might that complex be represented in a text in language-system. Here all allowed ways are candidates for consideration, including some in which the representation is not a neat package occurring between any particular pair of internal boundaries.

Efforts to mechanize searching are perhaps most closely related to mechanized translation in the case of certain approaches for which the following two characteristics hold: first, a machine-readable version of the full unaltered text of a document is prescribed, with the intention of mechanizing all subsequent processing involved in searching; second, the full text or some equivalent transformed version of it is to be accessible in the course of actual searching. Where the second characteristic applies, we can say that that document will be "completely" represented in the searchable store. If, on the other hand, only an abstract, an extract, a set of index entries, or some other partial surrogate of the document is to be accessible in the course of searching, we can say that the document will be "incompletely" or "selectively" represented. Selective representation thus encompasses what we customarily call "subject indexing", regardless of its style. The distinction between complete and incomplete (or selective) representation appears to be a useful one in the characterization of systems and approaches. It should be noted that the production and use of a machine-readable version of full text is not itself an indication that representation will be complete in the searchable store.

Against this background I should like now to discuss briefly some work on the development of mechanized searching systems which my colleagues and I are doing, with support from the National Science Foundation, in which the approach to representation is deliberately selective, and the selection process is rather highly systematized. One of our reasons for adopting this approach was our practical judgment that selective representation, albeit selective, is useful under certain real circumstances and will continue to be useful as far into the future as one cares to look.

In our current exploratory work, we do not assume or require that the selection process be fully mechanizable. Despite this fact, I think it is not impertinent to discuss the work on this occasion for a number of reasons, among them the following: first, the approach is one which does take off from linguistic data (not merely lexical data) obtained directly from the documents processed; second, linguistic problems encountered in processing and augmenting these data are problems also for searching systems based on completely represented texts; and third, selective representations of the sort generated in this approach could be a useful diagnostic tool in explorations of the effectiveness and comparative efficiency of prototype searching systems based on complete representation and also of related prototype systems in which selective representation (e. g. , extracts or index data) are produced mechanically from machine-readable texts.

In our conception of searching systems, we distinguish between stores for searching, wherein actual documents are represented, and auxiliary stores, which in company with other tools set forth linguistic facts (and other facts) about the system. We also distinguish between selection and representation; but we exploit the representational scheme as a device for systematizing selection procedures.

Let us consider for a moment the gap between the complete text of some typical scientific article, as published in a journal, and one portion of that document which is normally present, its descriptive title. To select that document from a large collection on the basis of its substantive relevance to a given topic, a practicing scientist often needs more than the title (which cues a little), and less than the text (which in one sense tells all). Our approach to selecting what is to be searchable is to prescribe that a descriptive title be used for what it is worth, and to prescribe, further, how it shall be elaborated and how it shall be supplemented by other title-like expressions on the basis of the text.

In its concepts and methods, the approach is not purely linguistic but rather logico-linguistic, in a wide sense of both logic (i. e. , formal logic) and linguistics. The title is first analyzed as a linguistic entity; on the basis of this analysis it is then reformulated by a kind of paraphrasing which is regulated by a "normal" grammar, related to functional logic. Elaborations of the title and supplemental title-like expressions are also formulated in the normal grammar.

Consider as an example a document entitled "Gamma rays from neutron inelastic scattering". In our jargon, one prescription for elaborating a normalized version of this title might read as follows: "For a function term present in major occurrence at the 'inmost' level of title structure (in our example, the term 'scattering') supply maximally specific proper argument terms for all roles of its basic role pattern." From the text of the document, we would arrive thereby at an exclusive disjunction of sixteen proper argument terms (beryllium, boron, carbon, etc.). The title thus elaborated would provide index data equivalent to twenty or so conventional index entries, not counting entries that could result from valid substitution of other terms, for example, more generic terms.

The schemes we are developing for normalized representation are intended to be capable of preserving those components of informative meaning which are conveyed syntactically in phrases subjected to normalization. For us, the effort has considerable scientific interest. But it is fair to ask under what practical conditions such preservation is useful or worthwhile. Without approaching the question in a general way, I can mention one real situation where utility is observable. In Chemical Abstracts, the number of documents abstracted and indexed annually exceeds 100,000. The subject-index entries are relatively detailed. The vocabulary employed in the entries is, in general, that of professional chemists, and it is used with great care. What happens, then, if we ignore syntactical features of the entries and search only for the co-occurrence of pertinent words within an entry? We have conducted a number of experiments of this sort with somewhat revealing results. With respect to an intended search question, responses valid by word co-occurrence, but otherwise invalid, varied widely, in fact from 0 to 100%. For example, when it was desired to find entries indicating formation of aromatic compounds from cyclohexane and its derivatives, under the heading "Cyclohexane", 60% of the responses were invalid; and when it was desired to find entries concerning alpha rays from the actinide elements, under the heading "Alpha Rays", some 24% of the responses were invalid.

In developing indexing prescriptions we are reaching for a scope and depth of selection comparable to that realized in Chemical Abstracts. From samples we have studied, it now appears that

indexing by title and by direct elaboration of title produces index data equivalent to those found in a Chemical Abstracts subject index for about 50% of the documents represented. For the remainder, other prescriptions are required.

We have a growing collection of linguistic information resulting from our analyses of actual titles. One of the looming features of independent phrases with title function is the frequent occurrence of lengthy attributive endocentric noun-phrases. Strings of as many as six or more non-particle words may be present in constructions of this type. In our experimental work on normalization we paraphrase to make explicit many of the relations among participating constituents.

In attempting to normalize natural-language expressions we encounter, of course, many problems. One problem is that of determining where, and over what scope, condensation or telescoping occurs in the case of endocentric constructions.

Another problem is that of coping with roving modifiers. In our experimental work our decision has been to try to locate such modifiers at points of maximum precision in the normalized structure.

One of the most pervasive problems of normalization is that of compensating for discrepancies between the logical types identifiable in a particular independent phrase and the logical types present in a structural pattern generalized to comprehend a class of such phrases.

For the titles we have thus far analyzed, instances of multiple meaning for individual words have been relatively few, as compared with instances of structural ambiguity.

It is not feasible to describe in a few minutes the factors we are considering in designing normalizing schemes or the features that are now incorporated. About two months from now an informal report presenting some of this material should be available.

The feature of a normalizing scheme of particular interest to this Symposium is, I think, its utility as a device for organizing and for aiding the discovery of a portion of the relations of informative meaning that obtain for a given natural-language system. Recognition of meaning relations at word level and at stem level is only the beginning of semantic analysis. We hope that our work will be of some use in the larger effort to extend the scope of systematic semantic description.