STRUCTURE IDENTIFICATION IN AUTOMATIC TRANSLATION

Since most current work in automatic translation seems to be concerned with written (printed, rather than spoken languages, the problem of identifying significant structures automatically is solved at the elementary level: letters unlike phonemes, can be easily identified. Words in the input language are also readily identifiable, leading to the possibility of designing automatic dictionaries. In general, it can be safely assumed that any individual string of arbitrary length occurring in an input text can be identified on a lexical basis, e.g. by matching with an isomorphic string held in machine memory. It follows that if we are willing to store a large enough number of long enough strings automatic translation is a trivial theoretical problem. Since all current estimates of "large enough" and "long enough" are astronomical numbers, other methods based on recursive algorithms for operations with strings of length of the order of that of words, are necessary.

The simplest approach is to consider successive input words to be independent, and to define the translation of a string of words to be the string of the translation of these words ordered as the original words. This is the simplest type of automatic dictionary, which promises to be capable of providing rough translations adequate for a significant range of applications, The major drawback of this approach arises from frequent one-to-many correspondences between input and output words.

The solution of the multiple correspondence problem has been shown to be intimately connected with the more general problem of transforming the string of translations of individual words into a string conforming with the syntactic conventions of the target language. In either case, a method for identifying classes of word sequences seems essential. Lexical methods for sequence-class identification could be constructed$_c$ but would necessarily rely on the use of extensive storage facilities, and large memory search times, The use of morphological characteristics of input strings to identify word-classes,, or such sequence classes as subject and predicate, subordinate clauses, etc., does not seem to have received sufficient consideration,, The use of punctuation patterns, gap distributions between function words, affix patterns, and similar morphological characteristics of input strings to define sequence classes should be seriously investigated*

A. G. Oettinger
Computation Laboratory
Harvard University