A PROGRAM FOR MACHINE TRANSLATION USING A HIGH-CAPACITY STORE

Notes for discussion at the meetings of Oct. 16-19 preceding the International Conference on Mechanical Translation on Oct., 20, 1956 at the Massachusetts Institute of Technology, Cambridge, Mass,

by

W.Ryland Hill
Professor of Electrical Engineering
University of Washington
Seattle

Recent technological developments indicate the possibility of constructing high-speed storage devices with much greater capacity than those formerly available. This paper sets forth a program for using such a store for machine translation of languages. Rather than attempting to reduce the required storage capacity by stem and ending dissection or by using "idioglossaries", this program is based on the assumption of a storage capacity large enough to store all predictable word forms and many useful word groups.

Basic to the proposed system is the concept of "semantic unit" — an extension of the ordinary concept of "word". The symbols of the input text employed in the translation process include not only the normal alphabet but also the inter-word space, punctuation marks, and any other graphic distinctions available. Thus the input symbols comprise an extended alphabet, and a semantic unit may include the space symbol and punctuation marks as part of its symbol sequence. Consequently, a semantic unit is any symbol sequence of the input text which by reason of source-target semantics requires translation as a single unit. Semantic units are commonly words but they may be prefixes, suffices, combining forms, or even idiomatlc sequences.

The machine translator requires a mechanism to search and compare the entries in the store with the input text for the purpose of breaking down the input into

semantic units which are then translated and printed out. The only reason for including semantic units containing several words is to obtain a better translation than is possible by dealing with the constituent words. Hence the machine is designed to break the sentences into the largest possible semantic units found in the machine vocabulary.

Obtaining the largest possible semantic unit matching a portion of the input texts requires a logical ordering of the store entries and a progressive search system. Including punctuation and other symbols as part of the alphabet and adding the concept of "blank" or alphabetic zero permits this ordering to be easily obtained. The arrangement is somewhat similar to that of an ordinary dictionary.

In dividing the input text into semantic units the machine does not depend upon the space symbol to indicate a dividing point. For this reason the basic comparison process can break compound words into constituents and also deal with prefixes and suffixes as well as translate simple words and idiomatic phrases.

The need for idioglossaries can be obviated by including processing tags with each store entry. Each entry consists of a source language semantic unit, control symbols initiating the comparison and read out operations, the target language words comprising this translation, editing symbols to help the reader, and processing symbols to reduce the multiplicity of target equivalents. Although one entry may include multiple target equivalents, each of these includes processing tags corresponding to the fields of knowledge for which the particular equivalent is most probable. The machine is programmed to select only the target equivalent corresponding to the particular field of knowledge covered by the input text. Thus the complete store in effect includes the idioglossaries for many fields of knowledge. Because the number of possible translations for a given semantic unit of the source language is far less than the number of fields of knowledge, this arrangement requires much less storage than would separate idioglossaries.

The approach outlined here gives a translation by semantic units instead of a. simple word-for-word translation but does not necessitate any extensive use of logical processing requiring a digital computer. For this reason future additions to the program for the purpose of utilizing contextual relationships to improve the

translation can be handled with the addition of a moderate amount of logical processing which would not then be superimposed on an already complex logical system.