[*Mechanical Translation*, vol.3, no.2, November 1956; pp. 40-41]

Contextual Analysis in Word-for-word MT

Kenneth E. Harper, Slavic Department, University of California, Los Angeles

EXPERIMENTS with word-for-word MT of Russian scientific literature have given results which, except for such limited purposes as indexing, are far from satisfactory. The difficulty is not so much one of word order as of syntactic and semantic ambiguity of individual words. Regardless of the treatment of the problem of inflected forms, for example, it is impossible in the majority of instances to identify the grammatical case of Russian nouns. In addition to syntactic ambiguity, multiple equivalents must be assigned to a large percentage of words (to an estimated 45% of the running words in a physics text). The chief disadvantage of word-for-word MT, then, is its prolixity: the reader is confronted with a burdensome multiplicity of potential equivalents (syntactic and semantic) for several words in each sentence.

The chief cause of this ambiguity is the fact that each word is examined in isolation, as a discrete item. The human translator operates with the tremendous advantage of something called "context". Broadly speaking, context signifies environment: surrounding words, sentences, and even the subject area itself. Investigation shows that restricted contextual analysis, performed routinely, can resolve most of the problems of ambiguity. Remarkable clarification is attained even when the comparison of a given ambiguous word x is limited to the immediately contiguous word in the sentence (the pre-x or post-x word). Without attempting to rearrange the word order of the Russian sentence, one can obtain the following by comparison of each ambiguous word with the coded grammatical features or semantic class of contiguous words:

a) Syntactic clarification. The ambiguity of case forms in nouns can be reduced to an insignificant percentage, and proper English equivalents can be supplied in the form of English prepositions as demanded by the genitive, dative, and instrumental cases. Such prepositions can be withheld in translation when the requirements of Russian grammar demand it. Participles and adverbs which are indistinguishable in form from adjectives, can, be given the correct equivalent; the comparative degree of adjectives and adverbs can be adequately handled. In general, there are no serious problems of syntax which cannot be resolved by reference to the grammatical features of pre- or post-words.

b) Semantic clarification. The correct English equivalents of most of the "glue words" (especially prepositions and conjunctions) can be found only through contextual analysis. The programming of such analysis should be based on the observed behavior of these words in actual conditions. Thus, the meaning of the conjunction "i", which has at least four equivalents (and, but, also, even) can be pinpointed in more than 90% of all occurrences by simple reference to the grammatical category of contiguous words; the pronoun-adjective "ikh", meaning "(of) their" or "(of) them", can be similarly resolved. It should be stressed that completely unpredictable and unexpected relationships can be found between structural context and meaning, and that the barest kind of routine comparison results in a high (although not absolute) degree of accuracy in the determination of meaning.

Non-structural clarification of meaning takes several forms. In the first place, techniques of MT lexicography need to be developed, i.e., the science of choosing the best "cover-all" target language equivalent from a group of relatively synonymous equivalents, and the selection of equivalents based on observed behavior, rather than upon the evidence of a dictionary. (Thus, in the area of physics the Russian izmenenie may always be found to equate with "change", although Bray's technical dictionary lists nine fairly distinct meanings.) In effect, what is needed are true ideoglossaries, based on actual, rather than potential, behavior.

The application of contextual analysis offers great potentialities for semantic clarification. Operating again on the basis of observation, we can construct and code word classes which cause contiguous words to behave in a predictable manner. Thus, the preposition po has ten potential possible equivalents when followed by a noun in the dative case; by reference to predetermined noun classes we can reduce the number of choices to one in a given instance. The necessity of treating each new combination as an "idiom" is eliminated. It is also possible

to pinpoint the meaning of many nouns which are ambiguous even within an ideoglossary by reference to the class of the accompanying adjective, or to specified key words in the title or opening sentences of the text.

There is no question that the kind of study in syntax and semantics which can be realized with the aid of machine techniques will result in the discovery of usable principles of association, so vital in the operation of what is called "contextual analysis".