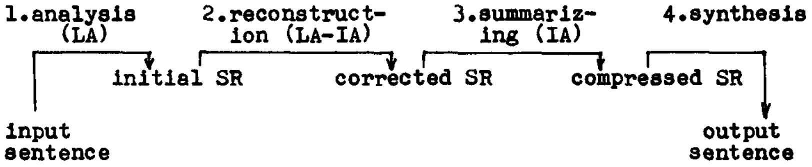
AUTOMATIC TRANSLATION THROUGH UNDERSTANDING AND SUMMARIZING

N. N. Leont´eva

Vsesojuznyj centr perevodov
Kržižanovskogo 14, 117 218 Moskva, USSR

French to Russian automatic translation system being developed in All-Union Centre of Translation is conceived as part of a multifunctional information processing system in the sense that it should be able to use approaches and methods proper to the information processing field, such as summarizing, abstracting, indexing, making inferences, etc. In such a system translation is realized through building text information representation (IR). The task requires two types of analysis: linguistic analysis (LA) and information analysis (IA) working in interaction, the latter being, in particular, able to refer to the automatic thesaurus. The ultimate aim of LA is the building of sentence semantic representation (SR). It is important that for each individual sentence its SR is constructed as a function of the IR of the whole text. (The current version of the system does not operate with the whole text but is limited for each sentence with its more or less immediate context.) Linguistic analysis calculates morphological structure for words, syntactic and semantic structures for sentences. Each of these structures is determined by the appropriate language realities; still remaining obscurities can be cleared only by referring to higher levels of analysis. SR built for an individual sentence without regard to other sentences´ SR´s is normally incomplete (deficient, ambiguous, incorrect, etc.). SR incompleteness is manifested by incompleteness of its units. The construction of text IR requires operations of comparison of different SR units as well as their comparison with thesaurus units. As a rule, incompleteness proper to SR´s is cleared only partially, which calls for some external measures to ensure a formally correct structure ready for the synthesis of the output text. The general scheme of the system functioning runs as follows:



Linguistic analysis contains a set of procedures aimed at creating initial SR´s where all cases of incompleteness are exposed. Reconstruction compares SR´s with each other and with the thesaurus and restaures the missing parts of SR´s. Summarizing means obtaining a kind of an abstract from which all obscure and incomplete parts are removed so that only essential information is available.

Information processing plays an important role in realisation of the scheme as the system translates only what it comprehends, thus the result may be called not a literal but a "digested" translation. The information model of automatic translation is based on the properties of the coherent text. One of the main properties is that pieces of information essential for the text are repeated there in many ways and by various linguistic means. IA aims at identifying such information and making it the basis for SR reconstruction. The level of "information noise" in the synthesized text is expected to be lower than in the classical approach to AT (sentence-to-sentence translation through syntactic structures). The degree of abstracting (summarizing) can vary depending on the purpose: the system can be oriented at getting a translation proper, a detailed or a brief abstract, a summary, or, finally, a search pattern. The effect of such reproductions of the input text with subsiding detality reminds of an echo which gradually loses almost all original features keeping the main pattern to the end:no degree of abstracting should affect the document main contents.

The system information orientation determines the choice of linguistic means of analysis, mainly, the structure and units of syntactic and semantic representations. Two principles can be formulated: "purity" of means at each level of analysis and possibilities of interaction between levels. The first principle makes it possible to use with maximum efficiency the laws specific to each level and to certify the formal correctness of the resulting structure. The second principle implies a kind of hierarchial organisation of grammar: if a unit of one level cannot be interpreted at a higher level, it can be "generalized" (a lexeme can be generalized to a semantic class, a labeled relation can be replaced by a more general or even an unlabeled relation). Building of a structure at each level comprises at least two stages: creation of the initial structure permitted to be incomplete and incorrect, and reconstruction of a more complete and correct structure after an interpretation of the initial structure by means of the higher level (or levels).

The division into levels is manifested not only by different means of analysis but also by different nature of units: nodes and relations. Nodes of syntactic representation are words (difference of lexical meanings is disregarded), nodes of semantic representation are lexical meanings, nodes of IR are notions having denotative status. Relations of syntactic structure are functional (from predicate to subject, form predicate to direct or indirect object, attributive relation, etc.). SR-relations are of semantic nature (cause, time, patient, etc.), IR relations are mainly the same but vary in their information value: some appear inside a notion and are devaluated, others connect separate notions and acquire denotative status.

Units of translation are represented by units of IR having an explicite inner structure and liable to translation either as a whole or by parts. They are formed in the course of both linguistic and information analyses.