R. G. Piotrowski - I. V. Palibina

AUTOMATIC PATTERN RECOGNITION APPLIED TO SEMANTIC PROBLEMS

The idea of computer pattern recognition as applied to semantics has the same basic principles as the theory of the perception, which represents a sufficiently generalized model for that part of the brain whose function is classification, insofar as classification presupposes similarity between elements (or groups of elements) of the same class and difference between elements (or groups of elements) of different classes.

In this respect automatic semantic pattern recognition (ASPR) consists in the following. An unknown text is analyzed by computer. At the first stage of the text treatment its subject matter is identified at the word-form level. The second stage results in indicative abstracting at the phrase (word-group) level.

The theory of semantic pattern recognition offers a number of concepts with their structural and functional particularities. First of all, these are concepts of

1) semantic space, as a certain semantic zone of the language;

2) semantic regions, as certain semantic subzones of this space;

3) distinguishing factors, those semantic elements in whose terms the semantic space is being formed;

4) pattern alphabet – a list of the region names and finally the concept of

5) semantic pattern itself.

The semantic pattern is regarded to be the name of a particular region of the semantic space wherein information structure on phenomena of the outer world is reflected. This information structure is represented as an ordered set consisting of objects (as extra-linguistic reality) and its relations. Any information on the phenomena of objective reality could be rendered by means of combinations of objects together with their relations. Any finite set of objects combined with the relations between them produces situation. So, properly speaking, the semantic pattern is a situation mould, *sui generis*, transferred to the na-

tural language. A real semantic space which is represented by a sublanguage is infinite and reflects all possible situations in the given sublanguage that has a concept system of its own. This concept system is reflected and fixed in texts as a whole complex of semantic relations along with *signifiants* of individuals. If the semantic space manifests itself as a specific sublanguage then its regions can naturally be represented by separate branches of knowledge, divisions and subdivisions of this sublanguage and even by a single descriptor. A list of the region names, into which the semantic space falls, makes an alphabet of patterns. Ideally, the pattern alphabet agrees with UDC entries.

The actual semantic space referred to, e.g. texts of any scientific and technical sublanguage is under study, to be exact, a finite model for this semantic space as word-form (or segment) frequency dictionary, obtained as a result of automatic text processing. By this means there is a model *a posteriori* whose structure follows naturally from the properties, particularities and semantic parameters of the source material. Frequency dictionaries representing a certain semantic zone of a scientific and technical sublanguage along with subject frequency microdictionaries, as semantic subzones of this sublanguage give a true lexical picture and reflect the specific content system for a given sublanguage. A word-form (segment) frequency dictionary of the texts under investigation serves as a basis for selecting semantic distinguishing factors, which are represented by the meaning of certain word-forms, words (key words) and segments (key segments). The construction of recognizing systems is a matter of some difficulty especially in the field of man-machine communications. Recognition is regarded to be the whole complex of processes in the system, including the process of learning. In forming a recognizing automaton we « rough-hew » the « internal » world of the computer, a functional model of high nervous activity, our attention being focused on experimental criteria only, that is « consciousness » is considered from a functional point of view. As we stated above, the recognition process falls into learning and the application of information obtained in learning, i.e. extrapolation of regularities of linguistic material, that has been registered into the storage to the source text in its processing. The structure of recognizing automaton should reflect the information nature of that branch of knowledge (specific sublanguage) on whose treatment it is oriented. It involves a variety of reference regions (*étalons*) of the semantic space of the sublanguage under treatment. These reference regions are represented by generalized patterns of sublanguage divisions which are being formed on the basis of distin-

guishing semantic factors obtained after statistico-distributional analysis of a great body of scientific and technical texts in English and Russian has been carried out. The reference region manifests as a certain model for a specific branch of knowledge with its characteristic features and particularities as it is represented in the designer's brain. The more comprehensive and deeper insight of the recognizing automaton designer into this branch of knowledge, the better the reference region serves the purpose of recognition. By the end of learning as many reference regions must be recorded in the storage as we can bring out analysing the texts under study. Each of these reference regions involves a number of distinguishing semantic factors that provide descriptions of the semantic space regions; the semantic factor selection follows the principle of establishing a strict line of demarcation between the regions in making their descriptions; either distinguishing semantic factors are available in a region of semantic space and not found in all the other regions, or they can possibly be found there.

Two possible situations arise from checking a semantic pattern of an unknown source text against reference regions recorded into the storage in the course of teaching computer.

1) Deterministic situation occurs when a description of a source text pattern agrees ($p = 1$) or doesn't ($p = 0$) with a reference region, it means, that the source text belongs to that scientific and technical branch of knowledge, whose model this particular reference region represents.

2) Probabilistic situation occurs when a partial agreement between semantic patterns of a source text and reference region description occurs. The coincidence probability takes all the values in the interval from 0 to 1.

Both deterministic and probabilistic models of semantic pattern recognition function at different levels such as lexical, morphologo-syntactical, semantico-syntactical.

Of the working semantic recognition systems which have been developed until now one can name the bilingual automatic dictionary (deterministic model at lexical level), the bilingual dictionary of segments (deterministic model at morphologo-syntactical level), multilingual thesaurus of local syntactical meanings (deterministic model at semantico-syntactical level), semantic, pattern recognition system for indexing, annotating and abstracting (both deterministic and probabilistic models at the integrated semantico-morphologo-syntactical level). In addition, a bilingual thesaurus development is under way.

In obtaining the information needed for the efficient performance of an automatic system for semantic pattern recognition we used the statistico-distributional technique combined with automatic text treatment. It offers a higher degree of formalization of the process and also of objectivity. Each of these factors is of prime importance if the problem of ASPR is to be worked out and realized on a commercial scale.