SovietRxiv

Soviet-era science, translated into English

Reports of the Academy of Sciences of the USSR

1960

SovietRxiv

View the original and related papers at https://sovietrxiv.org/items/ru-196001.98715

Source: Math-Net.Ru and CyberLeninka. Machine translation. Verify with the original.

SovietRxiv

Abstract

Full Text

Reports of the Academy of Sciences of the USSR
1960. Volume 132, No. 5

CYBERNETICS AND CONTROL THEORY

T. M. NIKOLAEVA

ON THE STRUCTURE OF AN ALGORITHM FOR INDEPENDENT GRAMMATICAL ANALYSIS OF THE RUSSIAN LANGUAGE

(Presented by Academician S. A. Lebedev, 19 VI 1959)

1. In machine translation, the text being analyzed may be correlated either directly with the translating text in another language, or with some grammatical system of relations.

The first experiments in machine translation followed the first path; at present, however, most research is being conducted by correlating the text with some system of relations. Such a system of relations may be either a universal grammar, or the grammar of the translating language, or the grammar of the analyzed language.

The approach to constructing an algorithm of analysis set forth in the present communication is based on the last principle. This means that the features produced in the course of analysis indicate the relation of the units of the text to the grammatical system of relations of the analyzed language—in the present case, Russian.

2. It is advisable to distinguish between the principle of constructing algorithms of analysis and the principles of describing the grammatical system of relations. The purpose of an analysis algorithm should be not to describe the grammatical system, but to determine how the units of the analyzed text correlate with this system. Naturally, in compiling such an algorithm it is desirable to find the shortest and most rational path toward solving this problem. This shortest path will not necessarily correspond to the usual sequence of sections in a grammatical description of a language.

3. Each element of the text (a word or a combination of words) must receive a definite number of features. A list of all these features is appended to the described algorithm for independent grammatical analysis of the Russian language, developed at the Institute of Precision Mechanics and Computer Engineering of the Academy of Sciences of the USSR. The main problem in the rational construction of the algorithm is such an arrangement of

its parts as will preserve the proper sequence in identifying features. This means that if some feature can be reliably inferred for a particular unit of the text on the basis of data already available, then this must be done immediately upon receiving these data. Consequently, the determination of such a feature is not postponed until the stage when all elements of the text belonging to that class can receive the corresponding features. For example, the features of case and number for the form " " can be obtained in the very first part of the algorithm (when the word is found in the dictionary), whereas for the form " "these features can be determined at a much later stage (after the syntactic segmentation of the sentence has been carried out).

4. Such an approach will reduce to a minimum the number of cases in which some form is assigned a provisional feature that, in the course of further analysis, may be replaced by another. The assignment of such provisional features seems inadvisable not only because it increases the number of commands executed by the machine, but also because it requires repeated reference to the same elements of the text.

In the present algorithm, however, it has not been possible to avoid completely such provisional features. These include, in particular, the feature , produced at the initial stage for forms of the type " ," in which the pri-

the sign of case can be produced at later stages. It proved necessary to introduce precisely this conditional feature because the indication contained in this feature of the potential possibility that a form with this feature is in the nominative case may be sufficient for breaking up a complex sentence into simple ones.

5. The place of specific rules in the general system of the algorithm is determined by the fact that some features can be derived from a combination of others. For example, in order to determine to which part of speech forms of the type " " belong—which may be an adverb, a short form of an adjective, or a special predicative form (in a construction of the type " ")—one must know whether there is a verbal form of the predicate in the sentence, as well as forms that can function only as the subject. Therefore the "verb" scheme must be placed before the operation of the group of rules that distinguish forms of the type " ."

6. First of all, those reliable features must be identified which may be needed in order to derive equally reliable features of other elements of the text. This explains, in particular, why in the present algorithm reliable features of nouns in the nominative-case form are first determined, as well as features of a verbal and any other predicate, since this is necessary for segmenting the sentence and for distinguishing different kinds of homonymy. It was not considered expedient to place the easily determined features of a number of oblique cases with nonhomonymous inflections at the beginning of the operation of the algorithm' s rules, since they are not needed for determining the features of other elements of the text.

7. The entire analysis algorithm is divided into two parts, separated by the schemes "sentence segmentation." In the first part, work proceeds with individual word forms without any reference whatsoever to the surrounding words; in the second part, work on determining the required features proceeds with the use of information about other words in the sentence, for which it proved necessary to segment the sentence, clearing it of various kinds of introductory constructions. Therefore elements belonging to one and the same part of speech are analyzed in different parts of the algorithm: thus, indeclinable nouns and homonymous case forms are analyzed in the second part. By the same principle, the scheme "adjective, part 2," where the syntactic relations of adjectives are analyzed, is placed between the two schemes "noun, part 2" and "noun, part 3," since in order to determine the connections of an adjective it is necessary to have information about the features of the noun, while adjectives which in this scheme have received the feature "noun" pass on to the next stage, to the scheme "noun, part 3," where such nouns are analyzed.

8. The ultimate goal of the analysis is to determine syntactic relations, which may be common with the translating language and therefore connect the analysis algorithm with the rules of synthesis. Syntactic features must be obtained by all significant parts of speech. They are obtained at different stages of the algorithm' s operation. The first stage of the algorithm' s operation consists in searching the sentence for any kind of predicate; the second, in searching for the subject. The third stage is the determination of syntactic links within a simple sentence; the fourth, the determination of syntactic links within a complex sentence. All these four stages follow one from another. In accordance with this, all elements of the text that have not received syntactic features earlier are directed to the penultimate scheme, "analysis of syntactic relations, part 1," where they receive the necessary features. The last part of the algorithm is the scheme "analysis of syntactic relations, part 2," where syntactic links are determined not for individual words, but for entire constructions and turns of phrase that correlate with certain parts of the main clause or with this clause as a whole.

Received
8 VI 1959

Note: Figure translations are in progress. See original paper for figures.

Source: Math-Net.Ru and CyberLeninka. Machine translation. Verify with the original.