RITA - AN EXPERIMENTAL MAN-COMPUTER SYSTEM ON A NATURAL LANGUAGE BASIS

A.P. Ershov
Computing Center. USSR Academy of Sciences
Siberian Branch, Novosibirsk

I.A. Mel'ohuk
Institute of Linguistics, USSR Academy
of Sciences, Moscow

A.3. Narlniany
Computing Center, USSR Academy of Sciences
Siberian Branch, Novosibirsk

Abstract

The report presents an experimental (Pictorial) Representation - Information - Text - Author system intended for work with texts in natural language and simple geometrical compositions. The general principles of the system operation, its architecture, and basic problems are discussed.

Introduction

RITA is an experimental software Picture<->Text system intended for operation within the bounds of a most simple "world" of graphic compositions constructed from a small number of elementary geometrical figures. The system is supposed to be capable of:

- constructing a graphic composition according to a given textual description.
- composing a sufficiently adequate and natural description for any geometrical composition.

It was agreed that both graphic composition and textual descriptions should be, at the opening stage of the project realization, as simple as possible.

COMPOSITION: One, two, or three circles located on a square screen. The proportions of the circles and their location on the screen, botl absolute and relative to each other arbitrary.

DESCRIPTION: One Russian sentence of a maximal number of 45 words, containing no comparisons or metaphors. To make up for the primitive geometrical "world", a reasonably large margin is suggested for the syntactic structure of descriptions.

Principles of System Realization

The system consists of two independently operating parts: 1) Language processor or L-processor and 2) Composition processor or C-processor, both intended for analysis as well as synthesis.

The two parts are linked together through an intexmediate level of the Semantic Presentation of Information (SemP) , see Fig. 1.

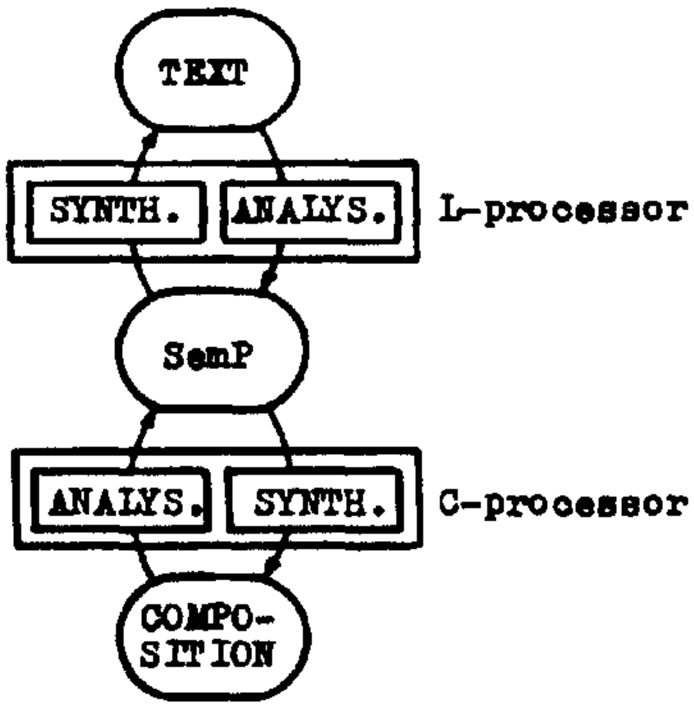

Figure 1

SemP, as a matter of fact, is an invariant corresponding, on the one hand,to a sufficiently broad class of synonymous texts, and, on the other hand, to a certain number of compositions; for any of these, every text of the class is an acceptable description. Apart from this , SemP is constructed as a formal object feasible for computer software representation. SemP acts as an intermediary language in Picture<=>SemP<=>Text transformations. SemP, in the RITA system, is a directed graph, its vertices marked with symbols of elementary (that is. within the frames of the system in question) "senses"*, i.e. "predicates'' and "objects" and arcs labelled with symbols of relations between the predicates and their arguments (note that predicate arguments may be predicates or objects, whereas vertices marked with object symbols are always terminal ones*).

The transformations corresponding to the phases of processor operation as indicated by pointers in Pig. 1 are not unique in a general case (with the exception of the Text *> SemP phase). It should be noted

The system conception is based on a general ideology of linguistic Text <=* Sense models presented in [1].

that:

1) the representation of different descriptions by one and the same SemP implies that these descriptions correspond to one and the same set of compositions;

2) the representation of a certain composition by more than one SemP does not imply that sets of compositions correspond'ing to each of these SemP are the same but, rather, that this composition belongs to the intersection of these sets;

3) since the world we confine ourselves to is extremely scanty and simple, sense units that need further semantic decomposition in a more general context may be used as elementary ones.

Consider a very simple example (see Fig. 2).

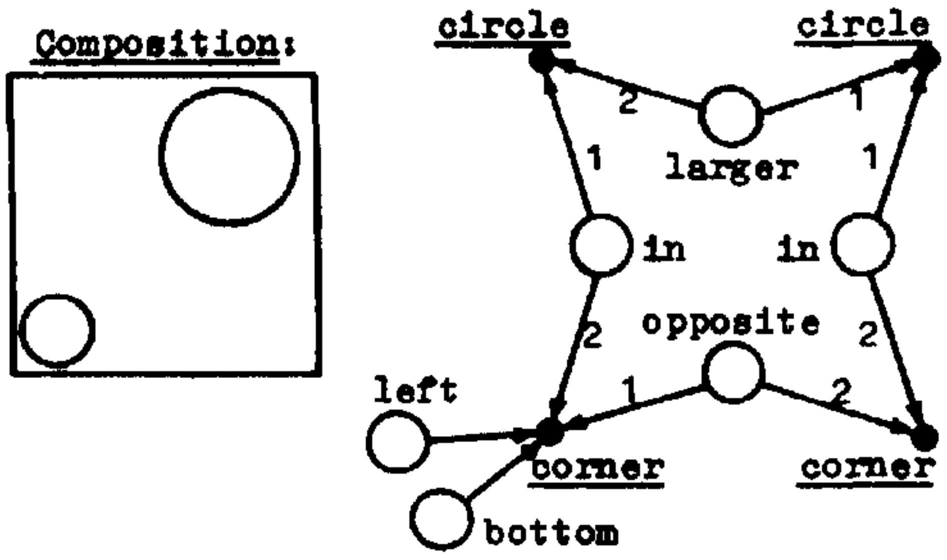

Figura 2

Text corresponding to SemP (I).

(1) A larger circle is located in the top right corner, and a smaller circle in the bottom left one.

(2) Of two circles on the screen,the smaller one is in the left corner on the bottom and the one that is large is in the opposite corner.

Texts corresponding to SemP (II) (Fig. 3).

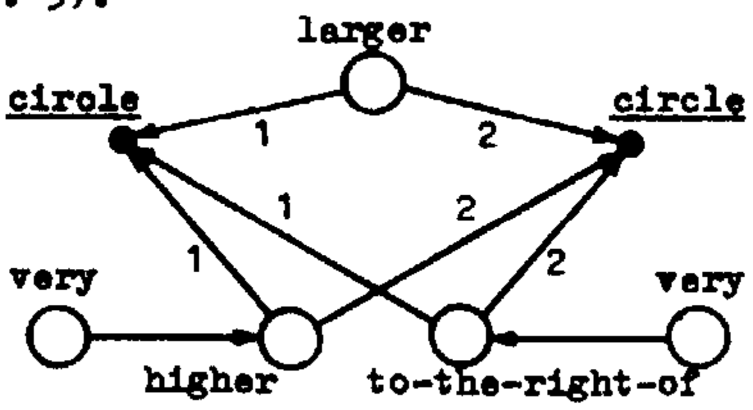

Figure 3

(3) There are two circles on the screen, the smaller one a deal lower that the larger one and on the left of the latter.

(4) Two circles are situated on the screen,one a lot higher than the other and to the right of it, that on top being bigger than the lower one.

System Architecture

A. The L-processor includes 3 main blocks. ANALYSIS, SYNTHESIS and DICTIONARY.

While in the analysis mode, the It-processor executes the standard routines 1 of natural text analysis:

1. Morphological analysis
2. Surface syntactic analysis
3. Deep syntactic analysis
4. Transition to SemP.

At the opening stage of system realization it was agreed to avoid the morphological analysis and include all necessary forms of words in the DICTIONARY. Thus , the sequence of the analysis mode stages is as follows:

Every word of the input sentence processed by the DICTIONARY is replaced by its entry form plus certain number of grammar characteristics (i.e. the so-called deep morphological representation of the word, or the word's DMR). The output string of the DMR's is an input for the ANALYSIS block. The surface analysis transforms the string into a surface syntactic tree (SSS). At the deep analysis phase , the SSS is made into deep syntactic structure tree (DSS). At the last stage of ANALYSIS, the DSS tree is transformed into a SemP.

The SYNTHESIS carries out the reverse function, i.e. transforming the SemP into a sentence of the output description.

The SYNTHESIS standard stages are:

1) Transition from SemP to a deep syntactic structure (DSS)
2) Transition from DSS to SSS
3) Transition from SSS to DMR string
4) Morphological Synthesis.

DICTIONARY: A preliminary study of more than 100 detailed textual descriptions of arbitrary compositions from a chosen class has shown that a dictionary of about 200 entries is sufficient for the current stage of the system realisation.

Since the morphological level is,for the time being, excluded from the L-processor (see above), the dictionary may contain as many as 2000 entries including all paradigms of words.

B. The C-prooessor consists of 4 main blocks (see Pig. 4);

1. Computational Model of Composition
2. Analysis-Synthesis of Predicates (ASP)
3. Predicate Filter
4. Analysis-Synthesis of SemP (ASSemP)

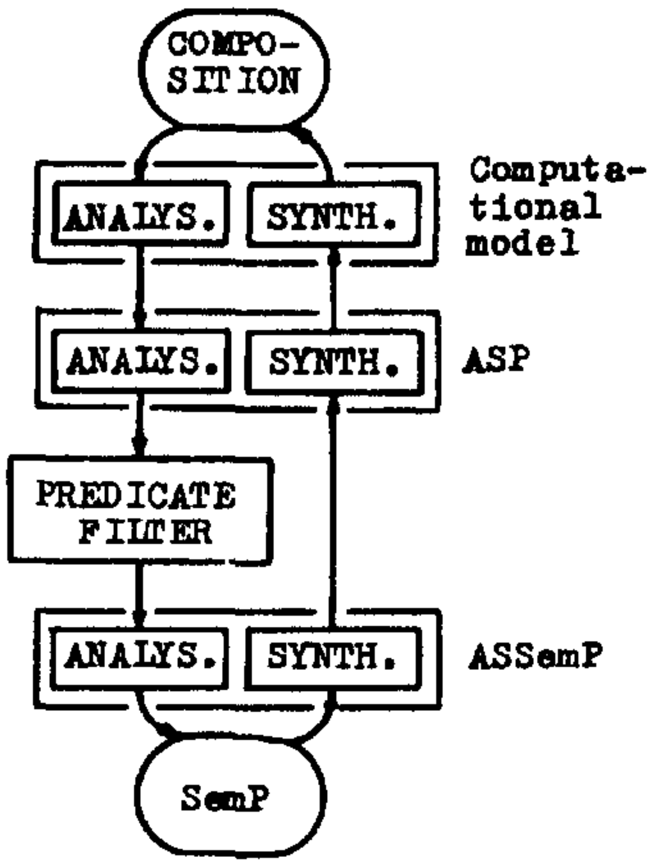

Figure 4

1) The computational model [5] of the composition uses a number of simple relatione defined on the set of elementary numerical characteristics of an image (e.g., the "distance between the 1st circle and the screen bottom edge", "radii ratio of the 2nd and 3rd circles", etc.) and statements concerning geometrical properties of image elements (e.g. "the second circle contacts the principal diagonal", "the 1st and 2nd circles intersect", etc.), the statements can be either "true" or "false".

Proceeding from given values of certain characteristics and statements, the model calculates values of other characteristics and statements, connected with the given ones through a system of "computable" relations. For example, in the Picture <=>SemP transforation, the basic characteristics are the circle and its center coordinates.

2) Analysis-Synthesis of Predicates sets a corresponding between values of statements and composition numerical characteristics and a certain set of elementary predicates which are SemP elements.

3) The predicate filter operates at the Picture=> SemP transition stage only. Prom the predicate set obtained at the previous stage the filter with the help of some heuristics selects a "representative" subject whose predicates have to be used as raw material in the SemP construction.

4) ASSemP: in the analysis mode,complex predicates are constructed from the chosen ones. These complex predicates together with some of the elementary predicates which have not been used before are made up into a SemP which must be "sufficiently adequate" and "minimally redundant". In the synthesis mode the input SemP is decomposed into predicates, complex predicates are further decomposed into elementary ones. The set of elementary predicates obtained is passed to the input of the synthesis module of the ASP block.

Discussion

The progress of large software systems requires a practical solution of the MAN-COMPUTER dialogue on a human language basis. Theoretical and experimental study of both linguistic and software components of a natural language dialogue systems is one of the principal aims of the RITA project. We also wish to confirm or correct our initial hypotheses as to what functional blocks of the L-processor can be recommended as a standard "preprocessor" for a more or less broad class of computer systems.

The preprocessor of any computer system capable of "understanding" a natural language includes three main components: dictionary, grammar (syntax + morphology), and semantics are defined by the system orientation, grammar being the only system-independent constituent of the preprocessor.

We believe that the text processing from a sentence to the Deep Syntactic Structure (DSS) level may be introduced as a universal block of the preprocessor. This block also defines the formats of the syntactic section of a dictionary entry, which can thus be made standard.

Besides the universal block the preprocessor must contain an interface block of the DSS SemP transformation. The interface realisation is entirely dependent on the SemP representation methods which, in their turn, must fit into the user system. It is evident that, for the standard part of the preprocessor changeable interface blocks should be supplied, each of them oriented on some specific class of computer systems (automatic management,retrieval systems, data banks, etc.).

The SemP language of interface determines the other (semantic) part of a dictionary entry. So, the computer system (or class of system) determines both the glossary of the preprocessor dictionary and the structure of the semantic part of The entries.

The L-processor constructed within the framework of the RITA system will serve as an experimental test of the reported ideology: first, a certain prototype of the universal part of the preprocessor is made up to the standards attainable at this stage, second, the operation of this universal part of the preprocessor the complex of the system as a whole.

As we wish to concentrate our main efforts on this problem the rest of the system is as much "lightened" as possible, its "world" being extremely limited, the SemP level simplified, a built-in learning mechanism renounced, etc.

However, the propect purposes are not restricted by the problems directly associated with the L-processor. In the course of the C-processor elaboration a number of problems also arise that require special investigation.

To grant an example, consider the concept "the circle is in the (screan)corner". Moving from text to image, we shall necessarily have to pass over from a formal expression of this oonoept on the SemP level to its interpretation in a most simple but "real" world of compositions. What sort of interpretations do we mean? Let us correlate each with an interpretative function (I-function) the values of which vary from the "firm YES" to the "firm NO" depending on values of certain parameters. For instance, with the world maximally simplified (the effect of other composition elements neglected, the corner and screen sizes (fixed), the I-function in our case depends on the three parameters! the two coordinates of the circle center and its radius. To make the I-function useful* we have to develop techniques of their generation and computer presentation. It offers no particular difficulties in the case of an I-function of a few variables; for functions of many variables, though, the problem is far from trivial. Most probably, the only possibility we have is to try to find approximation methods for I-functions, i.e. ultimately, to replace a complex concept with a super-position of simpler ones. The efforts to elaborate a sufficiently adequate procedure of such replacement can go in a few different directions:

1) "Explanatory dictionary": a restricted base set of "elementary" concepts is selected. For each non-elementary concepts and therefore is in itself an explanation (or "definition") of a non-elementary concept through elementary ones. The evident advantages of the approach are that

(1) the I-functions are forced into a "lower level" where they can be handled without particular difficulties and that

(ii) it enables deep synonymous paraphrasing on SemP level.

The disadvantage of the approach is that we do not dispose of an objective rocedure of building-up "explanations" although we can imagine an objective procedure of their verification). The drawback is the more troublesome as there are positively no proofs that the explanation technique is powerfull (or, rather, natural) enough to provide a "close" approximation for complex I-functlons of many variables*

2) Analytical approximation: the reconstruction, with analytic methods, of a complex I-function from partial data,e.g. individual cross-sections* projections, etc. We can also mention the approximation of I-functions of many variables with methods close to those of pattern recognition (when the situation is described as a set of values of simple concepts).

3) Research in techniques of a "natural" foxmation of complex concepts from simple ones, close to that in psychic and psychobiologioal mechanisms*

A combined strategy may prove to be the most officlent one: an explanatory dictionary up to a definite level, then methods of handling I-functions directly* The search for such a strategy and specific methods of its realization is also one of the main problems of the RITA project*

References

1. И.А.Мельчук. Опыт лингвистической теории моделей типа СМЫСЛ-ТЕКСТ, М.,I974.

2. И.А.Мельчук, А.С.Нариньяни. РИТА - экспериментальная система РИСУНОК-ИНФОРМАЦИЯ-ТЕКСТ, Доклад на УП Симпозиуме по кибернетике, Тбилиси, I974.

3. А.С.Нариньяни. Взаимодействие с машинными системами на естественном языке, Сборник "Системное и теоретическое программирование - 74", Новосибирск, I974.

4. A.P.Yershov. One View of Man-Machine Interaction. Journal of the ACM, vol. 12, No 3 (July, 1965).

5. Э.Х.Тыугу. Решение задач на вычислительных моделях, Журнал вычислительной математики и математической физики.