A RULE-BASED APPROACH TO EVALUATING IMPORTANCE IN DESCRIPTIVE TEXTS

Danilo Fum(•), Giovanni Guida(†), Carlo Tasso(‡)
Istituto di Matematica, Informatica e Sistemistica
Università di Udine
Udine, Italy

ABSTRACT

Importance evaluation is one of the most challenging problems in the field of text processing. In the paper we focus on the notion of importance from a computational standpoint, and we propose a procedural, rule-based approach to importance evaluation. This novel approach is supported by a prototype experimental system, called importance evaluator, that can deal with descriptive texts taken from computer science literature on operating systems. The evaluator relies on a set of importance rules that are used to assign importance values to the different parts of a text and to resolve or explain conflicting evaluations. The system utilizes world knowledge on the subject domain contained in an encyclopedia and takes into account a goal assigned by the user for specifying the pragmatic aspects of the understanding activity. The paper describes the role of the evaluator in the frame of a larger system for text summarization (SUSY); it illustrates its overall mode of operation, and discusses some meaningful examples.

1. INTRODUCTION

Text understanding has received increasing attention in recent years. A major problem in this area is that of importance evaluation: not all the components of a sufficiently large and structured piece of text are equally important for the reader, and humans are able to evaluate the relative importance of the parts of the texts they read.
This issue has been faced so far only in an indirect way in the literature on discourse structure (Kintsch and van Dijk, 1978; van Dijk and Kintsch, 1983), summarization (Lehnert, 1982 and 1984; Wilensky, 1982; Hahn and Reimer, 1984), and inference (Schank, 1979).
Moreover, several studies in the field of summarization (e.g.: Schank, 1979 and 1982; Lehnert, 1982 and 1984; Wilensky, 1982) have mostly been concerned with narrative texts (stories), and it is not at all obvious that the approaches which proved successful in this area could be applied to descriptive texts as well. Expository prose has its own specific features (Graesser, 1981), and it seems to require different understanding processes, different summarization skills, and different cognitive models (Lehnert, 1984). Work on the problem of understanding and summarizing expository prose is still at the very beginning (Hahn and Reimer, 1984).

In this paper we focus on the notion of importance from a computational standpoint, and we propose a rule-based approach to importance evaluation. This research is part of a larger project aimed at developing a system for understanding and summarizing descriptive texts (SUSY, a SUmmarizing SYstem), which is in progress at the University of Udine.
SUSY proposes an approach to descriptive text understanding and summarization (Fum, Guida, and Tasso, 1982, 1983, and 1984) in which the process of representing the meaning of a natural language text is split into three main tasks, namely: sentence understanding, structure capturing, and importance evaluation.

The sentence understanding phase works on the single sentences that constitute a given natural language text and maps them into a formal internal representation, called basic linear representation (BLR). The BLR is essentially a propositional language appropriately extended and completed to deal with the most relevant features of text representation, and fully worked out in a way suitable for computer implementation (Fum, Guida, and Tasso, 1984). The BLR representation of a text is constituted by a sequence of labeled propositions, each of them constituted by a predicate with instantiated arguments or representing an ISA relation between concepts. This phase includes the understanding of the literal meaning of each sentence in the text, the appropriate representation of time relations, and the treatment of quantification and reference.

The structure capturing phase works on the BLR and produces an augmented version of it, called extended linear representation (ELR). This phase focuses on two main points:

- inferring and expliciting the macro-structure of the text (van Dijk, 1977; Kintsch and van Dijk, 1978; van Dijk and Kintsch, 1983), that accounts for the conceptual connection (coherence) among sentences (Hobbs, 1982 and 1983);

(•) also with: Laboratorio di Psicologia E.E., Universita' di Trieste, Trieste, Italy
(†) also with: Milan Polytechnic Artificial Intelligence Project, Milano, Italy
(‡) also with: CISM - International Center for Mechanical Sciences, Udine, Italy

- recognizing and expliciting the rhetoric structure of the text, which explains how the flow of ideas and the arguments of the writer are organized and implemented in the text.

The importance evaluation phase operates on the ELR and attaches appropriate markers to its components in order to produce a new representation, called hierarchical propositional network (HPN). The HPN is a tree-like structure whose nodes, corresponding to concepts and propositions of the ELR, are assigned different importance values (integers) according to their relative importance in the text.

Once the HPN representation of a text has been produced, it is easy to prune the less relevant parts in order to obtain the representation of an appropriate summary to be eventually translated into natural language. These last phases (i.e., pruning and generation) are, for the moment, outside the scope of this research.

The purpose of this paper is to investigate in some detail the phase of importance evaluation, and to illustrate the results obtained in the design and experimentation of a prototype system that can produce from the ELR of a given text a reasonable HPN.
The paper is organized as follows: section two introduces the topic of importance evaluation and discusses some basic conceptual aspects, in section three the overall organization of the system is presented with particular attention to knowledge representation, section four illustrates some examples of importance evaluation, and section five concludes the paper.

2. EVALUATING IMPORTANCE

The topic of importance evaluation has been dealt with in recent years, although often only in a quite indirect way, by several authors and in many different contexts. A part of a text can be considered important in relation to other segments of the same text according to several criteria:

- it embodies knowledge necessary to understand other parts of the text (van Dijk, 1977; Kintsch and van Dijk, 1978);
- it is relevant to the topic of discourse (Lehnert 1982 and 1984);
- it is useful to clarify the relations that make discourse coherent (Hobbs, 1982);
- it relates to the topic-focus articulation (Hajičova' and Sgall, 1984);
- it refers to objects or relations in the subject domain that are judged to be important a-priori (Schank, 1979);
- it is unusual, new, or abnormal in the subject domain (Schank, 1979);
- it generates surprise (van Dijk and Kintsch, 1983);
- it is relevant to some specific reader's goal or need (Fum, Guida, and Tasso, 1982).

In practice, if we test these criteria on sample texts, they result sometimes complementary, sometimes partially overlapping, sometimes even conflicting. Moreover, different readers may judge differently the importance of the same text; on some parts a general consensus may be achieved, but the evaluation of other parts may be definitely subjective.
In fact, "important" means "specially relevant to some goal", and, whenever the goal with which a text is read changes, the parts of text which are to be considered important vary accordingly. Even if the goal of reading is only seldom considered explicitly by humans, still some goal is always implicitly assumed. Different readers (or the same reader in different moments) may have different goals, and conflicting judgments of importance may be due to the consideration of different goals, rather than to the application of different evaluation procedures.

The above investigation shows that importance is a really multifaceted concept which escapes a simple, explicit, algorithmic definition. A procedural, knowledge-based approach comprising a set of rules that can assign relative importance values to the different parts of a text and can resolve or explain conflicting evaluations seems more appropriate. Such an approach allows taking into account in a flexible and natural way the variety of knowledge sources and processing activities that are involved in importance evaluation. Moreover, it is expected to be well founded from a cognitive point of view (van Dijk and Kintsch, 1983; Anderson, 1976), as it allows close and transparent modeling of several processes that occur in human mind.

3. A COMPUTATIONAL APPROACH

Most of the ideas outlined in the previous section have been implemented in the design of a subsystem of SUSY, called the importance evaluator, that takes in input the ELR representation of a natural language text and the representation of a reader's goal and produces in output the corresponding HPN. The evaluator is implemented by a rule-based system (Davis and King, 1976) with a forward chaining control regime. Knowledge available to the evaluator comprises two parts: a rule base and an encyclopedia.

The rule base embodies expert knowledge necessary for importance evaluation. It is constituted by production rules, called importance rules, having the usual IF-THEN form. Rules can be

classified according to their competence, i.e. to the different types of knowledge utilized for evaluating importance. From this point of view, three classes of rules are considered:

- structural rules, which express the fact that some parts of the text can be judged important just by looking at their structure and organization, discarding their meaning;
- semantic rules, which can evaluate importance by specifically taking into account some specific structural features of the text that convey a definite meaning;
- encyclopedic rules, which can evaluate importance by comparing the meaning of the text with domain specific knowledge contained in the encyclopedia.

The IF-part of the rules contains conditions that are evaluated with respect to the current HPN (initially the ELR), and the THEN-part specifies either an importance evaluation or an action to be performed to further the analysis (e.g., a strategic choice, a criterion to solve conflicting evaluations, etc.).
The evaluation of importance contained in the THEN-part of a rule takes usually the form of an ordering relation (e.g., less, equal, etc.) among importance values of concepts or propositions of the ELR, or it specifies ranges of importance values (e.g., high, low, etc.). Thus, rules only assert relative importance of different parts of the text: a constraint propagation algorithm will eventually transform these relative evaluations into absolute importance values according to a given scale.

The encyclopedia is the second knowledge source employed by the evaluator and it contains domain specific knowledge. Encyclopedic knowledge is represented through a net of frames. Frames embody, in addition to a header, two kinds of slots:

- knowledge slots, that contain domain specific knowledge, represented in a form homogeneous with the propositional language of the ELR;
- reference slots, containing pointers to other frames that deal with related topics in the subject domain.

This organization allows easy implementation of a property inheritance mechanism.

We now illustrate the notion of goal which is of crucial importance for understanding the overall mode of operation of the evaluator. The goal is a chunk of variable knowledge, assigned by the user taking into account the pragmatic aspects of the understanding activity, that defines the motivations and objectives that are behind the reading process. The role of the goal is twofold:

- exerting control on the activation of importance rules that operate on the ELR;
- selective focusing, i.e. enabling the evaluator to choose from the encyclopedia the pieces of knowledge which are expected to be relevant to the current importance evaluation.

The use of the goal in selective focusing comprises two activities:

- validating matching between the current ELR and the knowledge contained in a frame header or knowledge slot (direct frame activation), or
- activating a new frame pointed at in a reference slot of a currently active frame (indirect frame activation).

Therefore, the encyclopedia does not contain any a-priory judgment about importance. Full responsibility of this activity is left to the evaluator, which can interpret the content of the encyclopedia frames according to the current goal and can use the extracted knowledge to support the rule-based evaluation process.

4. SAMPLE OPERATION OF THE EVALUATOR

The current prototype version of the evaluator operates on simple texts taken from scientific and technical computer science literature on operating systems. It includes about 40 importance rules and a small encyclopedia of about 30 frames. The goal has been assigned a very simple structure: it is a logical combination of key-terms, chosen in a predefinite set, that represent possible points of view a reader can take in analyzing a text.

In this section we will illustrate some of the most basic mechanisms of importance evaluation through a few examples.
Let us consider the following sample text:

"U-DOS is an operating system developed by Softproducts Ltd. in 1982. It has a modular organization and is suitable for real-time applications. U-DOS includes powerful tools for interactive processing and supports a sophisticated window management that makes it user friendly, i.e. easily usable by novices or untrained end-users. Easy operation is, in fact, the main reason of it widespread diffusion in the data processing market, especially among CAD/CAM users who appreciate its graphic utilities."

The ELR of this text results (for a complete description of the formalism refer to: Fum, Guida, and Tasso, 1984):

```
010 *OP-SYSTEM (U-DOS)
020 DEVELOP (SOFTPRODUCTS-LTD, U-DOS, T1)
```

```
        030 *PAST (T1)
        040 *YEAR-1982 (T1)
045 TIME-SPEC (40,20)
        050 HAVE (U-DOS, V1, P)
        060 *ORGANIZATION (V1)
        070 MODULAR (V1, P)
075 QUAL (70, 60)
        080 SUIT (U-DOS, VV2, P)
        090 *APPLICATION (VV2)
        100 REAL-TIME (VV2, P)
105 QUAL (100, 90)
        110 INCLUDE (U-DOS, VV3, P)
        120 *TOOL (VV3)
        130 POWERFUL (VV3, P)
135 QUAL (130, 120)
        140 APPROPRIATE-TO (VV3, V4, P)
145 QUAL (140, 120)
        150 *PROCESSING (V4)
        160 INTERACTIVE (V4, P)
165 QUAL (160, 150)
        170 SUPPORT (U-DOS, V5, P)
        180 *WINDOW-MANAGEMENT (V5)
        190 SOPHISTICATED (V5, P)
195 QUAL (190, 180)
        200 MAKE (V5, U-DOS, 210, P)
205 ENABLE (190, 210)
        210 USER-FRIENDLY (U-DOS, P)
215 CLARIFICATION (220, 210)
        220 OR (230, 260, P)
        230 EASILY (240, P)
        240 USE (VV6, U-DOS, P)
245 MOD (230, 240)
        250 *NOVICE (VV6)
        260 EASILY (270, P)
        270 USE (VV7, U-DOS, P)
275 MOD (260, 270)
        280 *END-USER (VV7)
        290 UNTRAINED (VV7, P)
295 QUAL (290, 280)
        300 REASON-FOR (310, 340)
305 RESULT (310, 340)
        310 EASILY (320, P)
        320 OPERATE (NIL, U-DOS)
325 MOD (310, 320)
        330 HAVE (U-DOS, V8, P)
        340 *DIFFUSION (V8)
        350 LARGE (V8, P)
355 QUAL (350, 340)
        360 IN (330, V9, P)
        370 *DATA-PROCESSING-MARKET (V9)
        380 AMONG (330, VV10, P)
385 SPECIFICATION (360, 380)
        390 *CAD/CAM-USER (VV10)
        400 APPRECIATE (VV10, VV11, P)
        410 *UTILITY (VV11)
        420 GRAPHIC (VV11, P)
425 QUAL (420, 410)
        430 HAVE (U-DOS, VV11, P)
```

The set of key-terms that can be used to specify the goal includes, among others: KNOW, BUY, and USE. We assume hereinafter the goal KNOW, i.e., we are particularly interested in knowing the main technical features of the U-DOS operating system. With such a goal, some pieces of the encyclopedia turn out to be relevant to the evaluation of our sample text, while others are discarded, as it will be illustrated below.

In order to analyze the text, the evaluator generates from the ELR, as a preliminary step, a new structure, called the cohesion graph, that explicitly shows all the references among propositions of the ELR. The cohesion graph is a bipartite graph whose nodes are constituted by concepts and propositions connected by three kinds of arcs:

- directed arcs connecting pairs of propositions (say from P to Q), which represent embedding of a proposition into another (Q in P);
- simple arcs, connecting a concept and a proposition, which indicate that the concept appears as an argument in the proposition;
- double directed arcs, connecting two concepts via a propositional node (say from A to B via P), which show that a concept enters as the argument of a proposition stating an ISA relation (P states that A ISA B).

A portion of the cohesion graph of our sample text is shown in Figure 1.

Structural rules can exploit the information provided by the cohesion graph in order to selectively capturing the importance of the different parts of the text. An example of a structural rule is:

Rule S4: Highly Referenced Concept

IF in the cohesion graph there is a concept C which is at least K-referenced
THEN assign C an importance value w(C) = high.

This rule guesses that a concept which is highly referenced in a text is probably important. In our example (where the parameter K is set equal to 3), the concept U-DOS is considered important as it is highly referenced.

Importance can be evaluated by chaining several rules. As an example, after rule S4 has been applied, the following rule can fire:

Rule M7: ISA Proposition

IF a proposition P represents an ISA relation
AND
the argument of P is a concept C with importance value w(C)
THEN assign P an importance value w(P) = w(C).

The rationale of this rule is that, if a concept is important, any proposition that states an ISA relation about that concept is important too. This allows, for example, considering proposition 10 (which states that U-DOS is an operating system) as important.
Rule M7 allows, moreover, the application

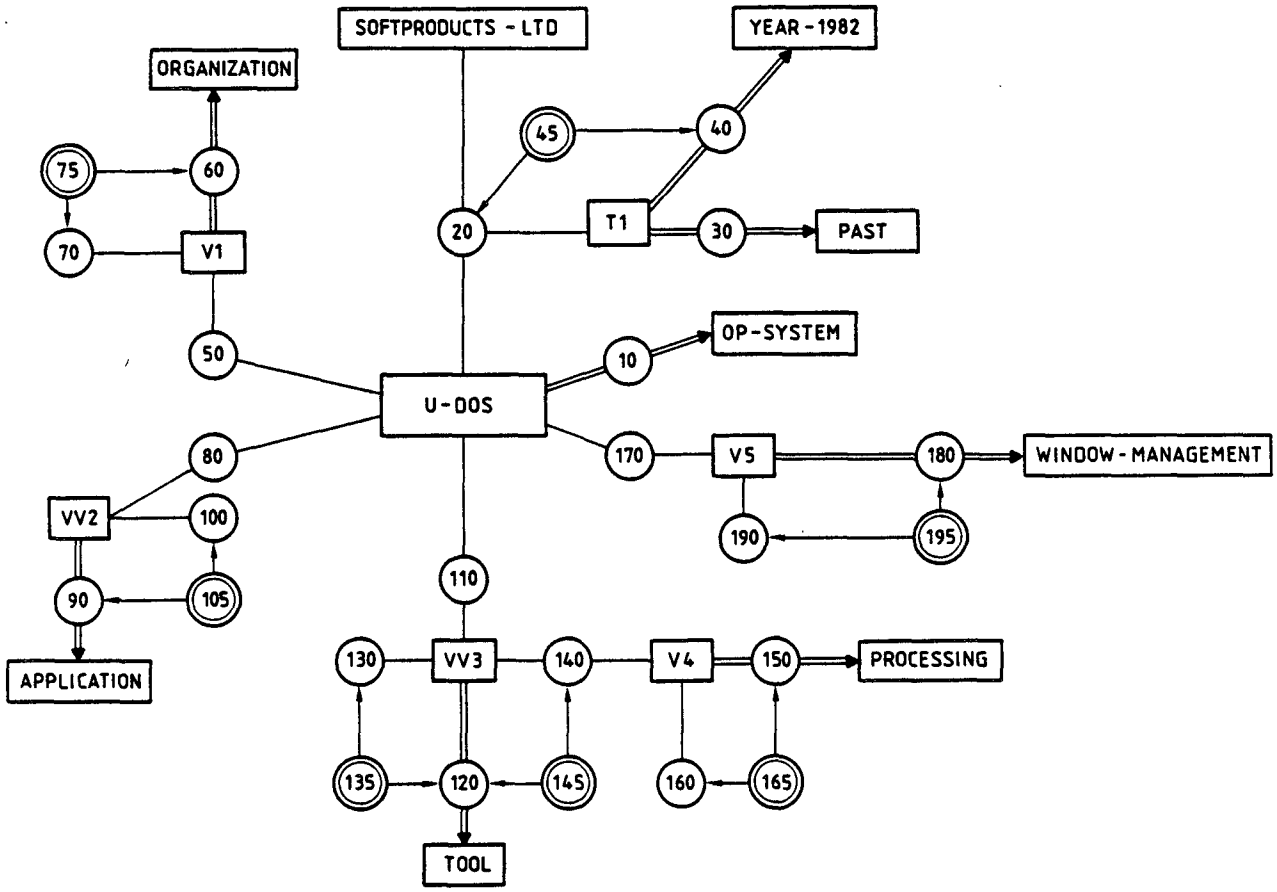

Fig. 1: (Partial) Cohesion Graph of the Sample Text

of the following rule:

Rule E6: ISA Frame Activation

IF a proposition P represents an ISA relation
AND
P has importance value w(P) > low
AND
the predicate of P is the header of a frame F in the encyclopedia
THEN activate F.

In our example, the fact that proposition 10 is important (w(P) = high) and that it represents an ISA relation allows the OPERATING-SYSTEM frame to be activated (see Figure 2, where a portion of the encyclopedia relevant to the current example is shown). Note that rule E6 does not directly state whether a proposition or a concept has to be considered important or not, but it specifies which frames are to be considered relevant in the current context.

Most evaluations are goal dependent and rely on a goal interpreter, able to evaluate a specific piece of ELR or a frame slot of the encyclopedia in order to determine its relevance to the current goal. The goal interpreter performs in such a way a complex matching, which allows implementation of selective focusing. Consider, for example, the following rule:

RULE E19: Goal-Dependent Frame Activation

IF the current goal matches a reference slot R of an active frame
THEN activate the frame whose header is pointed at by R.

Successive applications of this rule allow activation, starting from the OPERATING-SYSTEM frame, of the SOFTWARE-SYSTEM frame and, then, of the COMPLEX-SYSTEM frame (see Figure 2). At this point the following rule applies:

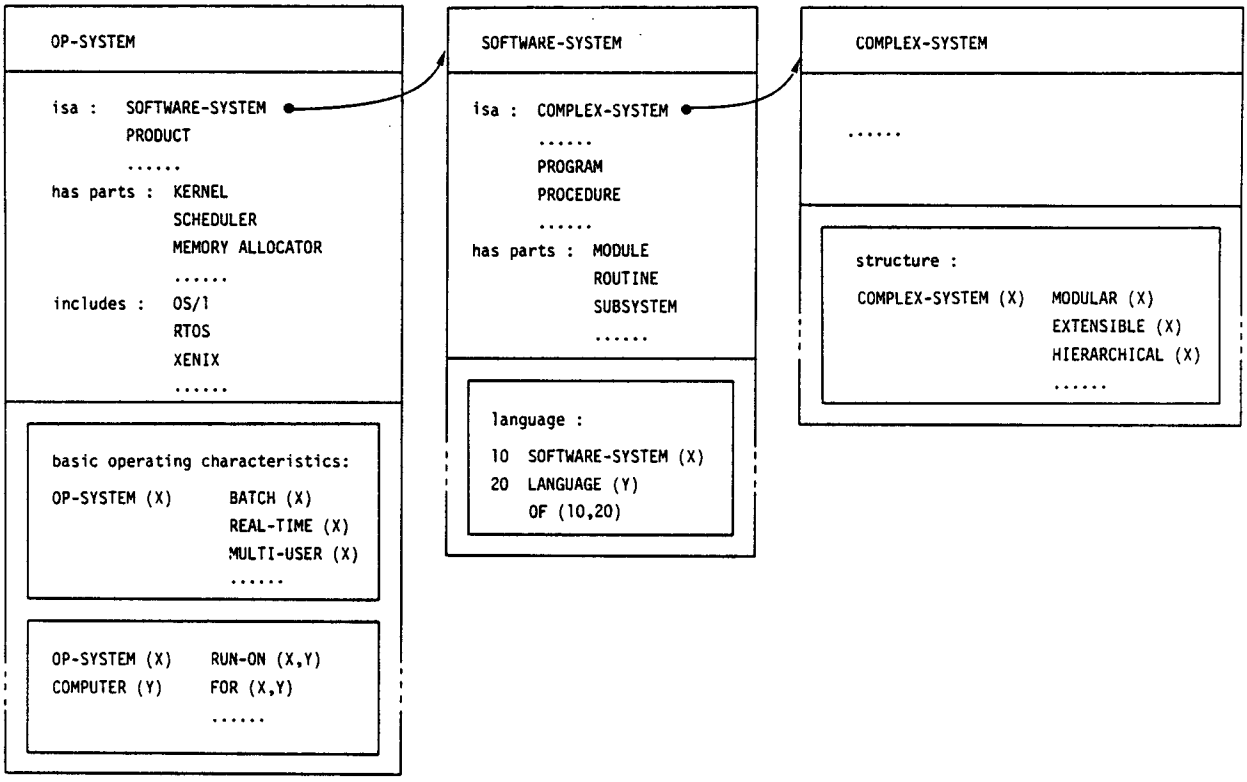

Fig. 2: Some Frames of the Encyclopedia

Rule E25: Goal Dependent Matching

IF a proposition P matches a pattern contained in a knowledge slot K of an active frame
AND
the current goal matches K
THEN assign P an importance value w(P) = high.

In our example, since (i) the COMPLEX-SYSTEM frame is active and proposition 70 of the ELR matches the pattern MODULAR (ORGANIZATION) of the "structure" slot of the frame, and (ii) the goal interpreter evaluates that the knowledge slot "structure" is relevant to the goal KNOW, then proposition 70 is considered important.

As a last example, we illustrate a rule that exploits knowledge concerning the macro-structure of the text:

Rule M9: Macro Clarification

IF if there exists a macro-proposition CLARIFICATION (P, Q)
THEN assign P and Q importance values such that w(P) < w(Q).

Rule M9 implements the idea that a proposition which is used to clarify another proposition (i.e., it paraphrases its content or explains the meaning of some of its terms) has to be considered less important than the proposition it clarifies. This rule can be applied, for example, in rating propositions 210 and 220, the latter resulting less important than the former.

5. CONCLUSION

The importance evaluator described in the previous sections is written in Franz Lisp and it is presently running in a prototype version on a SUN-2 workstation. Much experimental work is currently ongoing on this prototype in order to assess its operation, enlarge its knowledge base, and test its performance with a sufficiently large set of sample texts.

The major contribution of the work reported in the paper can be found in the novel proposed approach to importance evaluation that, according to the results so far achieved, proved to be viable and appropriate both from the cognitive and the computational points of view.

The research has disclosed several new directions for future work. Among these we mention:

- extending the importance rule base to cover the rhetoric and stylistic aspects of the text;
- introducing meta-rules to deal with the problems of rule activation scheduling, and of conflict resolution among rules;
- improving the goal matching techniques in order to implement a flexible mechanism for interpreting the content of encyclopedia frames according to the current goal;
- giving the evaluator the capability of changing the goal during the evaluation process, depending on the content of the processed text.

REFERENCES

1. Anderson J.R. (1976). Language, Memory, and Thought, Hillsdale, NJ: Lawrence Erlbaum.
2. Davis R. and King J. (1976). An Overview of Production Systems. In E.W.Elcock and D.Michie (Eds.), Machine Intelligence 8, New York, NY: Wiley, 300-332.
3. Fum D., Guida G., and Tasso C. (1982). Forward and Backward Reasoning in Automatic Abstracting. In J. Horecky (Ed.), COLING-82, Amsterdam, NL: North-Holland, 83-88.
4. Fum D., Guida G., and Tasso C. (1983). Capturing Importance in Natural Language Texts: An HPN-Based Approach. Proc. 2nd Int. Colloquium on the Interdisciplinary Study of the Semantics of Natural Language: Meaning and Lexicon, Kleve, FRG.
5. Fum D., Guida G., and Tasso C. (1984). A Propositional Language for Text Representation. In B.G. Bara and G. Guida (Eds.), Computational Models of Natural Language Processing, Amsterdam, NL: North-Holland, 121-163.
6. Graesser A.C. (1981). Prose Comprehension Beyond the Word. New York, NY: Springer-Verlag.
7. Hajičova' E. and Sgall P. (1984). From Topic and Focus of a Sentence to Linking in a Text. In B.G. Bara and G. Guida (Eds.), Computational Models of Natural Language Processing, Amsterdam, NL: North-Holland, 151-163.
8. Hahn U. and Reimer U. (1984). Computing Text Constituency: An Algorithmic Approach to the Generation of Text Graphs. In C.J. van Rijsbergen (Ed.), Research and Development in Information Retrieval, Cambridge, UK: Cambridge University Press, 343-368.
9. Hobbs J.R. (1982). Towards an Understanding of Coherence in Discourse. In W.G. Lehnert and M.H. Ringle (Eds.), Strategies for Natural Language Processing, Hillsdale, NJ: Lawrence Erlbaum, 223-244.
10. Kintsch W. and van Dijk T.A. (1978). Toward a Model of Text Comprehension. Psychological Review 85, 363-394.
11. Lehnert W.G. (1982). Plot Units: A Narrative Summarization Strategy. In W.G. Lehnert and M.H. Ringle (Eds.), Strategies for Natural Language Processing, Hillsdale, NJ: Lawrence Erlbaum, 375-414.
12. Lehnert W.G. (1984). Narrative Complexity Based on Summarization Algorithms. In B.G. Bara and G. Guida (Eds.), Computational Models of Natural Language Processing, Amsterdam, NL: North-Holland, 247-259.
13. Schank R.C. (1979). Interestingness: Controlling Inferences. Artificial Intelligence 12, 273-297.
14. Schank R.C. (1982). Reminding and Memory Organization: An Introduction to MOPs. In W.G. Lehnert and M.H. Ringle (Eds.), Strategies for Natural Language Processing, Hillsdale, NJ: Lawrence Erlbaum, 455-494.
15. van Dijk T.A. (1977). Semantic Macro Structures and Knowledge Frames in Discourse Comprehension. In M.A. Just and P.A. Carpenter (Eds.), Cognitive Processes in Comprehension, Hillsdale, NJ: Lawrence Erlbaum, 3-32.
16. van Dijk T.A. and Kintsch W. (1983). Strategies of Discourse Comprehension. New York, NY: Academic Press.
17. Wilensky R. (1982) Points: A Theory of the Structure of Stories in Memory. In W.G. Lehnert and M.H. Ringle (Eds.), Strategies for Natural Language Processing, Hillsdale, NJ: Lawrence Erlbaum, 345-374.