EVALUATING IMPORTANCE: A STEP TOWARDS TEXT SUMMARIZATION

Danilo Fun/ ', Giovanni Guida* \ Carlo Tasso^
Istituto di Matematica, Informatica e Sistemistica
Universita di Udinc
Udinc, Italy

ABSTRACT

The paper deals with the problem of evaluating importance of descriptive texts and proposes a procedural, rule-based approach which is implemented in a prototype experimental system operating in the specific domain of text summarization. Importance evaluation is performed through a set of rules which 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. In the paper some examples of the system operation are presented by following the evaluation of a small sample text.

INTRODUCTION

Understanding a written text is a complex process that exploits different capabilities including, among others, linguistic competence, common sense reasoning, and domain specific inference. This process can be divided into three main activities (Fum, Guida, and Tasso, 1984a):

1. understanding the literal meaning of every single sentence of the text (including reference, quantification, and time);

2. inferring and expliciting the macro-structure of the text that accounts for its global meaning and organization (including coherence, rhetoric, and stylistic relations);

3. evaluating the relative importance of the different conceptual units that constitute the text.

In recent years we have been working at developing a system (SUSY - a SUmmarizing SYstem) that can show some basic capabilities in performing the above mentioned activities in the specific domain of descriptive text summarization (Fum, Guida, and Tasso, 1982). In this paper we focus on the third activity only, namely importance evaluation, and we discuss the basic features and mode of operation of a module of the SUSY system devoted to this task.

The topic of importance evaluation has been dealt with, although often only in a quite indirect way, by several authors and in many different contexts. A conceptual unit of a text can be considered important in relation to other units according to several criteria that include, among others, relevance for explaining discourse coherence (Kintsch and van Dijk, 1978; Hobbs, 1982), relation to the topic (Lehnert, 1982) or topic-focus articulation (Hajičova' and Sgall, 1984) of the text, reference to semantically relevant concepts in the subject domain (Schank, 1979; van Dijk and Kintsch, 1983), relevance to a given goal (Fum, Guida, and Tasso, 1982).

In the paper we propose a new approach to importance evaluation (Fum, Guida, and Tasso, 1985) that integrates the above mentioned points of view into a unitary and flexible framework.

A RULE-BASED APPROACH

Two basic kinds of knowledge are involved in the process of evaluating importance in a text:

- linguistic knowledge, that makes possible to understand the meaning and structure of the text;

- world knowledge (including both common sense and domain specific knowledge) that is used for reasoning and inferencing.

In addition to this knowledge, we may assume that importance evaluation always relies on the (explicit or implicit) consideration of a goal. Furthermore, whenever the goal with which a text is read changes, the parts of the text that are judged important vary accordingly. Finally, knowledge about how to use linguistic knowledge, world knowledge, and goals in the process of importance evaluation, i.e. the criteria on which humans ground their judgment capabilities, has a crucial role too.

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

The complexity and expanse of knowledge involved in importance evaluation and the multifaceted nature of the processes that underly it, strongly suggest to resort to the powerful techniques offered by the rule-based system approach. In fact, the concept of importance seems to escape a simple, explicit, algorithmic definition. A procedural, knowledge-based approach comprising a set of rules that can assign relative importance values to the different conceptual units of a text seems more viable. This standpoint can supply the conceptual and computational tools needed for 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 (Anderson, 1976), as it allows close and transparent modeling of several processes that occur in human mind.

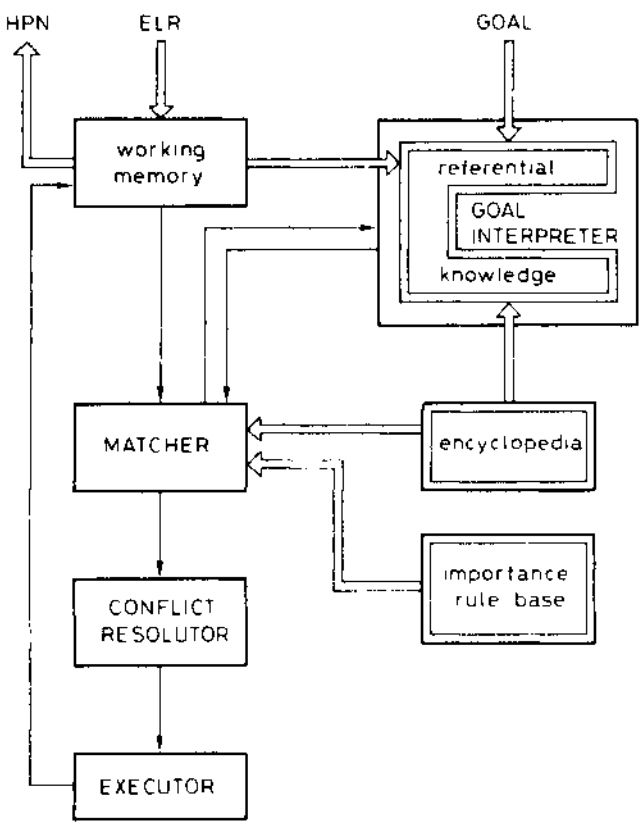

Figure 1 - Basic architecture of the evaluator.

On the basis of the above analysis, a prototype implementation of an experimental system, called importance evaluator, has been developed. This system is a Functional module of SUSY and concentrates on the importance evaluation task only. It receives in input the internal representation of a natural language text (supplied by another SUSY module, namely the parser) expressed in the ELR (Extended Linear Representation) formalism (Fum, Guida, and Tasso, 1984b), and produces in output a new representation called HPN (Hierarchical Propositional Network). In HPN integer importance values are assigned to the basic conceptual units of the ELR (concepts and propositions), in such a way as to account for the different importance of the constituents of the text. Moreover, the importance evaluator takes in input an explicit, declarative representation of a goal to be considered for its own activity.

The overall architecture of the evaluator is shown in Figure 1. It features a core rule-based structure with a forward-chaining control regime that includes a specialized module, namely the goal interpreter, devoted to make it fit the specific task of importance evaluation.
Two main knowledge bases are available to the evaluator:

1. the importance rule base, that contains knowledge (mostly of empiric nature) on the mechanisms that are supposed to be used by man in evaluating importance, expressed through IF-THEN production rules;

2. the encyclopedia, that contains specific world knowledge on the subject domain (mostly of structured, taxonomic, descriptive nature), represented through a network of frames.

The importance rule base includes several classes of rules:

- referential -structural (RS) rules that derive importance values from the structure of references among conceptual units of the text, taking into account, for example, concepts referenced in several propositions, propositions embedded into others, etc.;

" rhetoric-structural (TS) rules, that take into account the overall argumentational and stylistic organization of the text and derive importance relations from rhetoric-predicates of the ELR;

" structural-semantic (SS) rules, that rely on the analysis of some specific structural features of the text that have a definite semantic role, such as ISA relations, macro-predicates of the ELR, etc.;

~ semantic-encyclopedic (SE) rules, that refer to world knowledge contained in the encyclopedia concerning the specific subject domain dealt with by the text under consideration;

" explicit evaluation (EE) rules, that rely on explicit statements concerning importance evaluation that sometimes are purposely inserted in the text by the author in order to make reading and understanding easier;

" meta (MT) rules, that embody higher-level knowledge that concern reasoning about importance rules and that are used by the system mainly for solving conflicts between rule applications, i.e. for deciding which

rule to use first among several applicable ones, or which rule to trust among several conflicting ones.

The IF-part of a rule contains conditions that. *are* evaluated with respect to the current HPN (initially the ELR) contained in the working memory. The THFN-part specifies either an importance evaluation or an action to be performed to further the analysis (e.g., a strategic choice concerning rule activation, a criterion to solve conflicting evaluations, the activation of a frame of the encyclopedia, etc.).
The evaluation of importance contained in the THEN-part of a rule takes usually the form of an ordering relation among importance values of concepts or propositions of the ELR, or it specifies ranges of importance values (VERY HIGH, HIGH, MEDIUM, LOW, VERY LOW). Thus, rules only assert relative importance of different parts of the text: a constraint propagation al gorithm will eventually transform these relative evaluations into absolute importance values according to a given scale, after the evaluator has terminated its acti vity .

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 s 1 ot s :

- 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.

The operation of the evaluator obeys the basic recognize-act cycle shown below:

```
INITIALIZE working memory with ELR
CYCLE
   matcher activation
  MATCH the current working memory
      WITH the LHS of importance rules
          IF     the LHS of the considered rule
                 refers to the goal
          THEN   activate goal the interpreter
      DETFRMINE the set of all applicable rules
          IF     this set is empty
          THFN   EXIT CYCLF
   conflict resolution activation
  SELECT the rule to be applied next
   executor activation
  EXECUTE the RHS of the selected rule and update
          the working memory
END CYCLE
```

The above non-deterministic program deviates from the usual recognize-act cycle of a rule-based system because of the novel structure of the matcher that can invoke the goal interpreter whenever the goal is mentioned in a rule.

The goal is a chunk of variable knowledge expressed in a specific goal definition language (GPL). It is assigned by the user taking info account the pragmatic aspects of the understanding activity, and it 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 working memory, thus allowing implementation of evaluation mechanisms triggered by the current goal (goal-di rected evaluation);

  enabling the evaluator to choose from the encyclopedia the pieces of knowledge which *dr9.* expected to be relevant to the current importance evaluation activity, thus allowing the same pieces of world knowledge to be used differently in different situations according to the current goal (selecti ve focusing).

The specific way in which the goal influences operation of the matcher is determined by the goal i nterprefer. The motivation for having an interpreter for this activity can be found in the diversity between the language utilized by the user for stating a goal (the GDL) and the language in which the content of the text and the encyclopedia are represented (the ELR formalism) that does not allow a direct, and meaningful matching between pieces of knowledge expressed in these two languages. To this purpose the goal interpreter must have at its disposal an explicit representation of the semantic relationships existing between the worlds of the user goals and the knowledge on the subject domain. This additional knowledge is called referential knowledge as it can relate goals to specific" topics in thesubject domain. It takes the form of a network of conceptual chunks of knowledge (in the most usual cases, simple concepts) whose entry nodes represent items in the goal world and whose terminal nodes directly refer to frames of the encyclopedia or conceptual units of the ELR. The task of the goal interpreter is that of skillfully navigating in this network to find out the relevant relationships between the current goal and parts of the encyclopedia and HPN according to the conditions stated in the LHS of the rule being currently processed by the matcher.

The goal interpreter serves two basic functions that have a crucial role in the global architecture of the system. First, it allows to implement the encyclopedia without bothering of importance: no a-priori evaluation of importance is contained in it and full responsibility about importance evaluation is left to the rules. Second, it clearly separates the representation of world knowledge contained in the encyclopedia from the representation of goals, thus making any possible extension of the GDL easy and feasible without the need for restructuring knowledge in the encyclopedia.

THE IMPORTANCE EVALUATOR AT WORK

In this section we illustrate through the analysis of a sample text some of the most basic mechanisms of operation of the importance evaluator. The current prototype version of the evaluator (Fum, Guida, and Tasso, 1985) operates on scientific and technical computer science literature on operating systems. It contains about 40 importance rules and it comprises a small encyclopedia of about 30 frames. The goal definition language has been assigned a very simple structure: it allows to logically combine key-terms chosen in a predefinite vocabulary, that represent possible points of view of a reader (e.g., KNOW, USE, BUY, EVALUATE PERFORMANCE, etc.).

Let us consider the following fragment of a sample text:

"... An operating system is constituted by a set of programs which are used to monitor the execution of the user programs and the use of resources. One of the main reasons for utilizing operating systems is that they allow several processes to run at the same time. ..."

The ELR representation of the first sentence of this text is:

```
      010 CONSTITUTE (VV1, OP-SYSTEM, P)
      020 *PROGRAM (VV1)
      030 USE-FOR (NIL, VV1, 40, P)
  035 MACRO-GOAL (40, 30)
      040 MONITOR (VV1, 50, P)
      050 AND (60, 80)
      060 EXECUTE (NIL, VV2)
      070 *USER-PROGRAM (VV2)
      080 USE (NIL, VV3)
      090 *RESOURCE (VV3)
```

The importance evaluator usually tries to apply referential-structural rules first. An example of an RS rule is:

```
Rule RS4 - Highly Referenced Concept:
 IF   there is a concept X which is at least
      K-referenced
 THEN set w(X) = 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 5), the concept OP-SYSTEM is considered important as it is highly referenced in the ELR of the complete text.

After rule RS4 has been applied, the following structural-semantic rule can fire:

```
Rule SS5 - Definitional Predicate:
 IF   there is a proposition   P A(... X ...)
      such that  A ISA DEFINITIONAL,
      X is the 'definiendum'  of A,
      w(X) >= high
 THEN set w(P) = w(X).
```

Predicates of type DEFINITIONAL are used to describe the nature, properties, or essential qualities of a concept (e.g., DEFINE, EQUAL, CONSTITUTE, FORM, etc.). Rule SS5 conveys the idea that a proposition which defines a concept that is considered important inherits the importance value assigned to that concept. As a result of the application of this rule, proposition 10 receives the importance value w(10) = high.

After rule SS5 has been triggered, the following rule can fire:

```
Rule SS2 - ISA Proposition Extension:
 IF   there is a proposition P A(X)
      such that X ISA A,
      the  argument X  of  A  appears  in  another
      proposition  Q B(... X ...)  with importance
      value w(Q) >= high
 THEN set w(P) = w(Q).
```

This rule says that the proposition which specifies the type to which a concept contained in another important proposition belongs (ISA relation), is also important. In our example, proposition 20, which states that the variable VV1 represents a concept of type PROGRAM, receives the same importance value as proposition 10, i.e. w(20) = high.

It is interesting to note that the same result (i.e., w(20) = high) can be obtained, if we assume the current goal KNOW, through the successive application of rules SS6 and SS7. Let us examine both of them:

```
Rule SS6 - Goal-Directed Definitional Predicate:
 IF   there is a  proposition  P A(... X ... Y ...)
      such that  A ISA DEFINITIONAL,
      X is the  'definiendum'  of A,
      Y is the  'definiens'    of A,
      w(X) >= high,
      the current goal is KNOW
 THEN set w(Y) = w(X).

Rule SS7 - ISA Proposition:
 IF   there is a proposition P A(X)
      such that X ISA A,
      X  has  importance value w(X)
 THEN set w(P) = w(X).
```

Rule SS6 shows a case of goal-directed evaluation. The rule says that if the current goal is to know something about a concept which is important, the concept which is used to define it is to be considered important too. In our example, w(VV1) is set to high. This allows, in turn, application of rule SS7 which asserts that a proposition stating an ISA relation about an important concept is important too. Thus w(20) = high.

Propositions 30 and 40 receive a low importance value by the application of the following rule:

```
Rule SS12 - USE Propositional Inference:
 IF   there exist propositions   P A(NIL ... X ...)
                                 M MACRO-GOAL(P, Q)
                                 Q B(... X ...)
      such that A ISA USE,
                X is the 'object' of A,
                Q ISA ACT,
```

X is the 'agent' of B
THEN set w(P) = low, w(M) = low, w(Q) > low.

Rule SS12 implements a case of propositional inference (Graesser, 1981) that asserts that if something is used to do a certain action then it does that action, and only this is what matters.

Using the result previously obtained through rule RS4 (i.e., w(OP-SYSTEM) = high), we can now apply rule SE3:

Rule SE3 - Definitional Frame Activation:
IF there is a proposition P A(... X ...)
such that X ISA DEFINITIONAL,
X is the 'definiendum' of A,
w(X) >= high,
X is the header of a frame F
THEN activate F.

The idea on which SE3 is grounded is that, in order to understand a segment of text defining a concept which is judged important, it is necessary to have available the encyclopedic knowledge related to that concept, i.e., in our example, to activate the OP-SYSTEM frame. Note that rule SE3 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 to the current context. The mechanism of frame activation is commonly used in the operation of the evaluator. It models the well-known phenomenon of spreading activation that automatically occurs in human cognitive processes (Anderson, 1976).

In order to examine the use of the frame OP-SYSTEM above activated, let us introduce now the ELR of the second sentence of our sample text:

100 MAIN (VV4, P)
110 *REASON (VV4)
120 *VV4 (V5)
130 EQUAL (V5, 160, P)
140 REASON-FOR (160, 150, P)
145 MACRO-RESULT (160, 150)
150 USE (NIL, OP-SYSTEM)
160 ALLOW (OP-SYSTEM, 170, P)
170 RUN (VV6)
180 *PROCESS (VV6)
190 SEVERAL (VV6, P)
200 SIMULTANEOUSLY (170, P)

We can now apply the following rule:

Rule SE25 - Goal-Directed Matching:
IF there are propositions P1,...,Pn that match a pattern of a knowledge slot K of an active frame F,
the current goal matches K
THEN set w(P1) = ... = w(Pn) = high.

This rule states that if a piece of the ELR contained in the working memory matches the content of a knowledge slot of an active frame and the goal interpreter can relate the current goal to the content of this slot, then that piece of ELR is important. In our example, the knowledge slot TECHNICAL-FEATURES of the OP-SYSTEM frame includes, among others, the following fragment of knowledge:

10 RUN (VVx: PROCESS)
20 CONCURRENTLY (10)
30 DEFINE (MULTI-TASKING, 20)

Propositions 10 and 20 of the TECHNICAL-FEATURES knowledge slot match (indirectly through inferencing via ISA relations) propositions 170, 180, and 200 of the ELR, and, furthermore, the goal interpreter evaluates the slot TECHNICAL-FEATURES as relevant to the current goal KNOW. This yields w(170) = w(180) = w(200) = high.

REFERENCES

1. Anderson J.R. (1976). Language, Memory, and Thought, Hillsdale, NJ: Lawrence Erlbaum.
2. 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.
3. Fum D., Guida G., and Tasso C. (1984a). A Rule-Based Approach to Natural Language Text Representation and Comprehension. In R.Trappl (Ed.), Cybernetics and Systems Research 2, Amsterdam, NL: Elsevier Science, 727-732.
4. Fum D., Guida G., and Tasso C. (1984b). 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.
5. Fum D., Guida G., and Tasso C. (1985). A Rule-Based Approach to Evaluating Importance in Descriptive Texts. Proc. 2nd Conf. of the European Chapter of the Association for Computational Linguistics, Geneva, Switzerland.
6. Graesser A.C. (1981). Prose Comprehension Beyond the Word. New York, NY: Springer-Verlag.
7. Hajičová' 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. 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.
9. Kintsch W. and van Dijk T.A. (1978). Toward a Model of Text Comprehension. Psychological Review 85, 363-394.
10. 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.
11. Schank R.C. (1979). Interestingness: Controlling Inferences. Artificial Intelligence 12, 273-297.
12. van Dijk T.A. and Kintsch W. (1983). Strategies of Discourse Comprehension. New York, NY: Academic Press.