COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

NATURAL LANGUAGE ACCESS TO STRUCTURED TEXT

Jerry R. Hobbs, Donald E. Walker, and Robert A. Amsler
SRI International
Menlo Park, California 94025
U.S.A.

This paper discusses the problem of providing natural language access to textual material. We are developing a system that relates a request in English to specific passages in a document on the basis of correspondences between the logical representations of the information in the request and in the passages. In addition, we are developing procedures for automatically generating logical representations of text passages, directly from the text, by means of an analysis of the coherence structure of the passages.

INTRODUCTION

At SRI we are developing a system for natural language access to textual material. The system is to provide access to a textbook or other document of some importance, by returning relevant passages in response to a user's natural language request. Currently we are using the Hepatitis Knowledge Base, a compendium of current knowledge about hepatitis compiled by the National Library of Medicine, although the techniques we are devising are in no way particular to this document [cf. Walker, 1982]. The project has two phases. In the first, we are developing text access procedures for translating a user's request into an underlying logical form and, in order to locate the appropriate passages, matching the logical form with a Text Structure which expresses the structure of the document as a whole and summarizes the content of individual passages in terms of canonical predicates [Walker and Hobbs, 1981]. In the second, longer-term effort, we are developing procedures for automatically generating portions of the Text Structure directly from the text.

THE TEXT ACCESS COMPONENT

In the text access component, a user's request is translated into logical form by SRI's DIALOGIC system, described in another paper submitted to this conference [Grosz et al, 1982]. This logical expression is then turned over to the inferencing component DIANA [Hobbs, 1980], where various discourse problems are solved and a match with the Text Structure is sought.

As an illustration of this process, consider the following example query:

> During what period is immunoprophylaxis appropriate following exposure to type B hepatitis?

DIALOGIC translates the request into the following form:

```
DURING (APPROPRIATE (IMMUNOPROPHYLAXIS (I, X1, Y) |
                     FOLLOW (I, EXPOSE(X2, HEPATITIS-B))),
        ?X | PERIOD (?X) )
```

That is, during period ?X, the immunoprophylaxis I of X1 against Y, where I follows an exposure event of X2 to hepatitis B, is appropriate.

Two kinds of discourse problems are exemplified here. First, there is the problem of determining implicit arguments. We are not told explicitly what

immunoprophylaxis is against, only what exposure was to. We need to draw the inference that exposure to something is typically followed by immunoprophylaxis against it. This problem must be solved if we are to retrieve the proper passages on immunization against hepatitis B virus (HBV) rather than some other agent. Similarly, we are not told explicitly that the one who was exposed is the one who will receive immunoprophylaxis, that is, that X1 and X2 are the same individual.

The second discourse problem illustrated here is that of metonymy. One may talk about both exposure to HBV and exposure to type B hepatitis. In the first case we are talking about exposure to a virus, in the second exposure to a disease. The Text Structure is expressed in canonical predicates in a standardized form, and one of the standardizations is in the class of entities that can be the argument of a predicate. We must decide, for each predicate, the type of arguments it can take. For example, is one exposed to a virus or a disease? For various reasons, we have decided that one is exposed to a virus and not to a disease. Thus the inferencing procedures have to analyze the actual query into one involving exposure to the virus causing type B hepatitis, or to HBV. This coercion is done by accessing information in a knowledge base that "expose" requires a virus as its second argument, that type B hepatitis is caused by HBV, and that HBV is a virus.

In order to match the request with the Text Structure, DIANA needs to translate the original request into the canonical predicates in which the Text Structure is expressed. For example, since "immunoprophylaxis" is not one of the canonical predicates, we need to use the axiom

IMMUNOPROPHYLAXIS (i,p,v) iff IMMUNIZE(i, p, PROPHYLAXIS(v))

that is, i is an immunoprophylaxis event of p against v if and only if i is an immunization event of p for prophylaxis against v. The result is a translation into the canonical predicates "immunize" and "prophylaxis", which are used in the summaries of the relevant passages in the Text Structure.

GENERATING TEXT STRUCTURE

Our work on the automatic generation of the Text Structure is at a more preliminary stage. Automatic summarization is a central aspect of this effort. A certain amount of work has been done in artificial intelligence and psychology on the automatic construction of summaries, including work by Rumelhart [1975], Mandler and Johnson [1977], Schank and his colleagues [Schank et al., 1980], and Lehnert et al. [1981]. Most of this work has focused on narratives rather than expository discourse, however.

There are two principal techniques that we have brought to bear on the problem. The most important involves a coherence analysis of the paragraph, in a manner described in detail in Hobbs [1976, 1978] and similar to work by Longacre [1976] and Grimes [1975].

It can be argued that, in coherent discourse, one of a small number of coherence relations, such as parallel and elaboration, holds between successive segments of the text. The coherence relations can be defined in terms of the inferences that can be drawn from what is asserted by the segments being linked (called the assertions of the segments). Thus, very roughly, two sentences are parallel if their assertions make the same predications about similar entities.

These coherence relations allow one to build up a tree-like coherence structure for the whole text recursively, as follows: The coherence relations are defined between segments. A clause (perhaps elliptical) is a segment. When some coherence relation holds between two segments, the two together constitute a composed segment, which can itself be related to other segments of the text.

Since the coherence relations are defined in terms of the assertions of segments, we need to specify what the assertions of the composed segments are. For this purpose we use a number of heuristics. For example, if two sentences are

parallel, it is because the same predication is made about similar entities. Then the assertion of the composed segment makes that same predication about the superset to which the similar entities belong. Thus, every node in the coherence structure has an assertion associated with it. Very frequently the assertion associated with the top node of the coherence structure of a passage can function as the summary of the passage.

As an illustration of this technique, consider the following passage:

(P1) Blood probably contains the highest concentration of hepatitis B virus of any tissue except liver. Semen, vaginal secretions, and menstrual blood contain the agent and are infective. Saliva has lower concentrations than blood, and even hepatitis B surface antigen may be detectable in no more than half of infected individuals. Urine contains low concentrations at any given time.

After a grammatical analysis, the sentences in this passage can be aligned as in Figure 1.[1] Every clause considers some body material containing HBV in some concentration. They are thus linked by the parallel coherence relation, and the assertion (and the summary) of the passage is as follows:

CONTAIN (BODY-MATERIAL, HBV, CONCENTRATION)

Many paragraphs we have analyzed in this way turn out to have a parallel structure, and thus their summaries can often be constructed in a similar manner.

blood	contains	highest concentration	HBV
semen vaginal secretions menstrual blood	contain		agent
saliva	has	lower concentrations	
(saliva of) infected individuals	in	detectable ... no more than half	HBsAg
urine	contains	low concentrations	

Figure 1 Parallels in Passage (P1)

A second factor must also be taken into account in constructing the summarizations. In addition to containing summaries of individual passages, the Text Structure contains a representation of the hierarchical organization of the document as a whole, as well as other aspects of its overall structure. The place of an individual passage within the hierarchical organization constrains what can function as a summary of the passage. A summary must distinguish a passage from other passages at the same level in the hierarchy. Top-down considerations frequently lead us to refine a summary we arrive at solely by the bottom-up coherence analysis.

As an example, consider the following passage:

(P2) Generally blood donor quality is held high by avoiding commercial donors, persons with alcoholic cirrhosis, and those practicing illicit self-injection. Extremely careful selection of paid donors may provide safe blood sources in some instances.

---

[1] This diagram is similar to the formats developed by Sager and her colleagues [Sager, 1981].

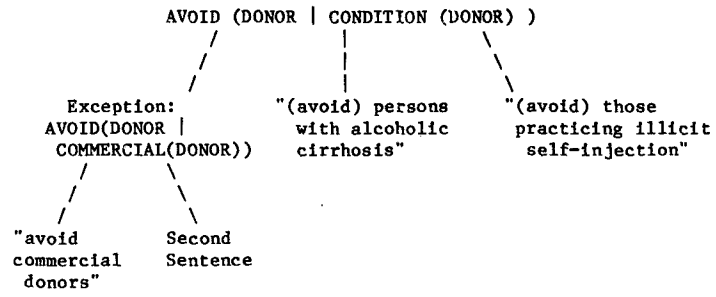

Figure 2 Coherence Structure of Paragraph (P2)

A coherence analysis results in the structure show in Figure 2. "Selection" contrasts with "avoiding," so we can say that the second sentence expresses an exception to the first conjunct of the first sentence. Because the second sentence is hedged very heavily, the assertion of the composed segment is the assertion of the initial conjunct of the first sentence--"avoid commercial donors." The three assertions of the first sentence stand in a parallel relation since they imply the same proposition about similar entities. They all imply (trivially) that certain classes of potential donors are to be avoided if blood quality is to be held high. Entities are similar if they share some common and reasonably specific property, that is, if they belong to some common and reasonably small superset. Our three classes of potential donors are similar in that they are all potential donors. The similarity would be stronger if there were some more specific property that characterized commercial donors, those with alcoholic cirrhosis, and illicit self-injectors, but there does not seem to be such a property. The most we can say seems to be that they are potential donors, and we arrive at the following assertion for the paragraph as a whole.

AVOID (DONOR | CONDITION (DONOR))

However, such a summary fails to distinguish this paragraph from its siblings in the hierarchical structure of the HKB as a whole. The nodes most immediately dominating this section in the hierarchy of the HKB correspond to sections about the quality of blood products under varying conditions, with respect to the risk of hepatitis in transfusion. There are two broad classes of conditions that are discussed, first, conditions characterizing the donor, and second, conditions characterizing the type of blood product. Among the conditions characterizing the donor are a history of hepatitis, recent transfusions, and positive results on serologic tests, as well as the conditions described in the example. Thus, the structure of the summaries in the paragraphs should be something like that shown in Figure 3.

It is therefore not sufficient for us to characterize the paragraph as being about avoiding potential donors exhibiting some condition. Thus, top-down considerations lead us to reject the summary we came up with solely by the bottom-up coherence analysis. We need something more specific, and the best we can do is simply to have a disjunction of properties as the condition characterizing the donors:

AVOID (DONOR | COMMERCIAL(DONOR) or CIRRHOSIS(DONOR)
or SELF-INJECTOR(DONOR))

```
QUALITY (BLOOD-PRODUCT)
   QUALITY (BLOOD-PRODUCT | CONDITION (DONOR) )
      [summary of our example]
      CONDITION = history of hepatitis
      CONDITION = recent transfusion
      CONDITION = positive serologic tests
      ....
   QUALITY (BLOOD-PRODUCT | TYPE (BLOOD-PRODUCT) )
      ....
```

Figure 3 Hierarchical Structure of Paragraph Summaries

CONCLUSION

While these methods for the automatic generation of summaries of expository text seem promising, difficult problems remain--including the problems of encoding and searching a very large knowledge base. In order to have practical milestone systems in the near term, we are working toward two scaled-down versions of the ultimate system. First, we are experimenting with using a pre-existing Text Structure to aid in the construction of the summaries of modifications of a passage. Second, rather than fully automatic generation of summaries, we are experimenting with ways that interaction with the author of a passage can aid in the task.

ACKNOWLEDGMENTS

This work has been supported by the National Library of Medicine under Grant 1-R01-LM03611.

REFERENCES

Grosz, B; Haas, N; Hobbs, J; Martin, P; Moore, R; Robinson, J; Rosenschein, S. 1982. "DIALOGIC, A Core Natural Language Processing System." COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics, Prague, Czechoslovakia.

Grimes, J. 1975. The Thread of Discourse. The Hague: Mouton.

Hobbs, JR. 1976. A Computational Approach to Discourse Analysis. Research Report 76-2, Department of Computer Sciences, City College, City University of New York (December 1976).

Hobbs, JR. 1978. Why Is Discourse Coherent? Technical Note 176, SRI International, Menlo Park, California (November 1978).

Hobbs, JR. 1980. Interpreting Natural Language Discourse. Final Report, National Science Foundation Research Grant No. MCS 78-07121 (July 1980).

Lehnert, WG; Black, JB; Reiser, BJ. 1981. "Summarizing Narratives." Proceedings of the Seventh International Joint Conference on Artificial Intelligence pp.184-189. New Haven, Connecticutt: Yale University.

Longacre, R. 1976. An Anatomy of Speech Notions. Ghent: The Peter de Ridder Press.

Mandler, J; Johnson, N. 1977. "Remembrance of Things Parsed: Story Structure and Recall." Cognitive Psychology 9:111-151.

Rumelhart, D. 1975. "Understanding and Summarizing Brief Stories." In: Basic Processing in Reading, Perception, and Comprehension, edited by D Laberge and S Samuels. Hillsdale, New Jersey: Lawrence Erlbaum.

Sager, N. 1981. Natural Language Information Processing: A Computer Grammar of English and Its Applications. Reading, Massachusetts: Addison-Wesley.

Schank, RC; Lebowitz, M; Birnbaum, L. 1980. "An Integrated Understander." American Journal of Computational Linguistics 6:13-30.

Walker, D. 1982. "Natural-Language-Access Systems and the Organization and Use of Information." COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics, Prague, Czechoslovakia.

Walker, DE; Hobbs, JR. 1981. "Natural Language Access to Medical Text." Proceedings of the Fifth Annual Symposium on Computer Applications in Medical Care pp 269-273. New York: IEEE.