GENERATING RELEVANT EXPLANATIONS: NATURAL LANGUAGE RESPONSES TO QUESTIONS ABOUT DATABASE STRUCTURE*

Kathleen R. McKeown
Department of Computer and Information Science
The Moore School
University of Pennsylvania
Philadelphia, Pa. 19104

ABSTRACT

The research described here is aimed at unresolved problems in both natural language generation and natural language interfaces to database systems. How relevant information is selected and then organized for the generation of responses to questions about database structure is examined. Due to limited space, this paper reports on only one method of explanation, called "compare and contrast". In particular, it describes a specific constraint on relevancy and organization that can be used for this response type.

I INTRODUCTION

Following Thompson [14], the process of generating natural language may be divided into two interacting phases: (1) determining the content, force, and shape of what is to be said (the "strategic component") and (2) transforming that message from an internal representation into English (the "tactical component"). The decisions made in the strategic component are the focal point of the current work. These decisions are being investigated through the development of a system for answering questions about database structure that require some type of explanation or description. This work, therefore, has two goals: (1) providing a facility that is lacking in many natural language interfaces to database systems, and (2) exercising theories about the nature of natural language generation. The system has been designed and implementation is in its beginning stages [12].

The decisions that the strategic component of a natural language generator must make are of two different types: decisions of a semantic/pragmatic nature and decisions that are structural in nature. Given a question, the strategic component must select only that information relevant to its answer (semantic/pragmatic decisions). What is selected must then be organized appropriately (structural decisions). These two types of decisions are the issues this work addresses. Not covered in this paper are the syntactic issues and problems of lexical choice that a tactical component must address.

Structural issues are important since the generation of text and not simply the generation of single sentences is being considered. A number of organizational principles that can be used for structuring expository text have been identified [12]. These are termed compare and contrast, top-down description, illustration through example, definition, bottom-up description, and analogy. In this paper, discussion is limited to compare and contrast and its effect on the organization and selection processes.

II THE APPLICATION

Current database systems, including those enhanced by a natural language interface (e.g. [6]), are, in most cases, limited in their responses to providing lists or tables of objects in the database.* Thus, allowable questions are those which place restrictions upon a class of objects occurring in the database. To ask these kinds of questions, a user must already know what kind of information is stored in the database and must be aware of how that information is structured.

The system whose design I am describing will answer questions about the structure and organization of the database (i.e. - meta-questions)**. The classes of meta-questions which will be accepted by the system include requests for definitions, requests for descriptions of information available in the database, questions about the differences between entity-classes, and questions about relations that hold between entities. Typical of such meta-questions are the following, taken from Malhotra [7]:

What kind of data do you have?
What do you know about unit cost?
What is the difference between
material cost and production cost?
What is production cost?

* This work was partially supported by NSF Grant MCS 79-08401 and an IBM Fellowship.

* Note that in some systems, the list (especially in cases where it consists of only one object) may be embedded in a sentence, or a table may be introduced by a sentence which has been generated by the system (e.g. - [4]).

** I am not addressing the problem of deciding whether the question is about structure or contents.

III KNOWLEDGE REPRESENTATION

In order for a system to answer meta-questions, it requires information beyond that normally encoded in a database schema. The knowledge base used in this system will be based on a logical database schema similar to that described by Mays [9]. It will be augmented by definitional information, specifying restrictions on class membership, and contingent information, specifying attribute-values which hold for all members of a single class. A generalization hierarchy, with mutual exclusion and exhaustion on sub-classes, will be used to provide further organization for the information. For more detail on the knowledge representation to be used, see [12].

IV SAMPLE QUESTIONS

Textual responses to meta-questions must be organized according to some principle in order to convey information appropriately. The compare and contrast principle is effective in answering questions that ask explicitly about the difference between entity-classes occurring in the database. (It is also effective in augmenting definitions but this would require a paper in itself.) In this paper, the following two questions will be used to illustrate how the strategic component operates:

(1) What is the difference between a part-time and a full-time student?
(2) What is the difference between a raven and a writing desk?

V SELECTION OF RELEVANT INFORMATION

Questions about the difference between entities require an assumption on the part of the speaker that there is some similarity between the items in question. This similarity must be determined before the ways in which the entities differ can be pointed out.

Entities can be contrasted along several different dimensions, all of which will not necessarily be required in a single response. These include:

attributes
super-classes
sub-classes
relations
related entities

For some entities, a comparison along the lines of one information type is more appropriate than along others. For example, comparing the attributes of part-time and full-time students (as in (A) below) can reasonably be part of an answer to question (1), but a comparison of the attributes of raven and writing desk yields a ludicrous answer to question (2) (see (B) below).

(A) A part-time student takes 2 or 3 courses/semester while a full-time student takes 3 or 4.

(B) A writing desk has 4 legs while a raven has only 2.

One factor influencing the type of information to be described is the "conceptual closeness" of the entities in question. The degree of closeness is indicated by the distance between the entity-classes in the knowledge base. Three features of the knowledge base are used in determining distance: the generalization hierarchy, database relationships, and definitional attributes. A test for closeness is made first via the generalization hierarchy and if that fails, then via relationships and definitional attributes.

A successful generalization hierarchy test indicates the highest degree of closeness. Usually, this will apply to questions about two sub-types of a common class, as in:

> What is the difference between production cost and material cost?
>
> What is the difference between a part-time and a full-time student?

In the generalization hierarchy, distance is determined by two factors: (1) the path between the entity-classes in question and the nearest common super-class; and (2) the generality of the common super-class (path between the common super-class and the root node of the hierarchy). The path is measured by considering its depth and breadth in the generalization hierarchy, as well as the reasons for the branches taken (provided by the definitional attributes). Entities are considered close in concept if path (1) is sufficiently short and path (2) sufficiently long. If the test succeeds, a discussion of the similarity in the hierarchical class structure of the entities, as well as a comparison of their distinguishing attributes, is appropriate.

Although the entities are not as close in concept if this test fails, some similarities may nevertheless exist between them (e.g. - consider the difference between a graduate student and a teacher). A discussion of similarities may be based on relationships both participate in (e.g. - teaching) or entities both are related to (e.g. - courses). In other cases, similarities may be based on definitional attributes which hold for both entities. For both cases, a discussion of the similarities should be augmented by a description of the difference in hierarchical class structure.

Entities that satisfy none of these tests are very different in concept, and a discussion of the class structure which separates them is informative. For example, for question (2) above, indicating that ravens belong to the class of animate objects, while writing desks are inanimate results in a better answer than a discussion of their attributes.

VI TEXT ORGANIZATION

There are several ways in which a text can be organized to achieve the compare and contrast orientation. One approach is to describe similarities between the entities in question, followed by differences. Alternatively, the response can be organized around the entities themselves; a discussion of the characterizing attributes of one entity may be followed by a discussion of the second. Finally, although the question may ask about the difference between entities, it may be impossible to compare them on any basis and the compare and contrast must be rejected.

The determination of the specific text outline is made by the structural processor of the strategic component. On the basis of the input question, the structural processor selects the organizing principle to be used (for the two sample questions, compare and contrast is selected). Then, on the basis of information available in the knowledge base, the decision is reevaluated and a commitment made to one of the outlines described above. Because of this reliance on semantic information to resolve structural problems, a high degree of interaction must exist between the structural processor and the processor which addresses semantic and pragmatic issues.

One type of semantic information which the structural processor uses in selecting an outline is, again, the distance between entity-classes in the knowledge base. For entities relatively close in concept, like the part-time and the full-time student, the text is organized by first presenting similarities and then differences. By first describing similarities, the response confirms the questioner´s initial assumption that the entities are similar and provides the basis for contrasting them. Two entities which are very different in concept can be described by presenting first a discussion of one, followed by a discussion of the other. Entities which cannot be described using the compare and contrast organization are those which have very little or no differences. For example, if one entity is a sub-concept of another, the two are essentially identical, and the compare and contrast organizing principle must be rejected and a new one selected.

VII STRATEGIC PROCESSING

Although dialogue facilities between the structural processor (STR) of the strategic component and the semantic/pragmatic processor (S&P) have not yet been implemented, the following hypothetical dialogue gives an idea of the intended result.

Question (1): What is the difference between a part-time and a full-time student?

STR: notes form of query and selects COMPARE AND CONTRAST

S&P: queries knowledge base:

```
     DISTANCE(part-time,full-time) -> very
     close (same immediate super-classes)
```

```
STR: retains COMPARE AND CONTRAST
     selects outline:
       SIMILARITIES
       DIFFERENCES:
          ATTRIBUTE-TYPE1
              .
              .
              .
          ATTRIBUTE-TYPEn
       CONSEQUENCES*
```

```
S&P: queries knowledge base  and fills in outline:

     SIMILARITIES
       super-classes(part-time,full-time)
           -> graduate student
       attribute/value(part-time,full-time)
           -> degree-sought = MS or PhD
     DIFFERENCES
       attribute/value(part-time,full-time)
           -> courses-required =
                part-time: 1 or 2/semester
                full-time: 3 or 4/semester
           -> source-of-income =
                part-time: full-time job
                full-time: unknown
     CONSEQUENCES none
```

STR: further organizational tasks, not described here, include determining paragraph breaks (see [12]). Here there is 1 paragraph.

The tactical component, with additional information from the strategic component, might translate this into:

> Both are graduate students going for a masters or Phd. A full-time student, however, takes 3 or 4 courses per semester, while a part-time student takes only 1 or 2 in addition to holding a full-time job.

After engaging in similar dialogue for question (2), the strategic component might produce outline (C) below, which the tactical component could translate as (D):

```
(C)  RAVEN FACTS:
       super-classes(raven) =
          raven E bird E animate object
     WRITING DESK FACTS:
       super-classes(writing desk)=
          writing desk E furniture E
          inanimate object
     CONSEQUENCES:
       bird and furniture incompatible
       2 different objects
```

(D) A raven is a bird and birds belong to the class of animate objects. A writing desk is

* CONSEQUENCES here involve only minimal inferences that can be made about the class structure.

a piece of furniture and furniture belongs to the class of inanimate objects. A bird can´t be a piece of furniture and a piece of furniture can´t be a bird since one is animate and the other isn´t. A raven and a writing desk therefore, are 2 very different things.

VIII RELATED RESEARCH IN GENERATION

Those working on generation have concentrated on the syntactic and lexical choice problems that are associated with the tactical component (for example, [10], [3], [13], [11]). Research on planning and generation ([1], [2]) comes closer to the problems I am addressing although it does not address the problem of relevancy and high-level text organization. Mann and Moore [8] deal with text organization for one particular domain in their generation system, but avoid the issue of relevancy. The selection of relevant information has been discussed by Hobbs and Robinson [5] who are interested in appropriate definitions.

IX CONCLUSIONS

The effects of a specific metric, the "conceptual closeness" of the items being compared, were shown on the organization and selection of relevant information for meta-question response generation. Other factors which influence the response, but were not discussed here include information about the user´s knowledge and the preceding discourse. Further research will attempt to identify specific constraints from these two sources which shape the response.

The research described here differs from previous work in generation in the following ways:

1. Previous work has concentrated on the problems in the tactical component of a generator. This work focusses on the strategic component: selecting and organizing relevant information for appropriate explanation.

2. While previous work has dealt, for the most part, with the generation of single sentences, here the emphasis is on the generation of multi-sentence strings.

When implemented, the application for generation will provide a facility for answering questions which the user of a database system has been shown to have about the structure of the database. In the process of describing or explaining structural properties of the database, theories about the nature of text structure and generation can be tested.

ACKNOWLEDGEMENTS

I would like to thank Dr. Aravind K. Joshi, Dr. Bonnie Webber, and Eric Mays for their detailed comments on the content and style of this paper.

REFERENCES

[1] Appelt, D. E. "Problem Solving Applied to Language Generation" In Proc. of the 18th Annual Meeting of the ACL. Philadelphia, Pa., June, 1980, pp. 59-63.

[2] Cohen, P. R. "On Knowing What to Say: Planning Speech Acts," Technical Report # 118, University of Toronto, Toronto, Canada, 1978.

[3] Goldman, N. "Conceptual Generation" In R. C. Schank (ed.), Conceptual Information Processing. North-Holland Publishing Co. Amsterdam, 1975.

[4] Grishman, R. "Response Generation in Question-Answering Systems" In Proc. of the 17th Annual Meeting of the ACL. La Jolla, Ca., August, 1979, pp. 99-101.

[5] Hobbs, J. R. and J. J. Robinson "Why Ask?", Technical Note 169, SRI International, Menlo Park, Ca., October 1978.

[6] Kaplan, S. J. "Cooperative Responses From a Portable Natural Language Data Base Query System", Ph.D. Dissertation, Computer and Information Science Department, University of Pennsylvania, Pa., 1979.

[7] Malhotra, A. "Design Criteria for a Knowledge-based English Language System for Management: an Experimental Analysis", MAC TR-146, MIT, Cambridge, Ma., 1975.

[8] Mann, W. C. and J. A. Moore "Computer as Author - Results and Prospects", ISI/RR-79-82, ISI, Marina del Rey, Ca., 1980.

[9] Mays, E. "Correcting Misconceptions About Database Structure" In Proc. of the Conference of the CSCSI. Victoria, British Columbia, Canada, May 1980, pp. 123-128.

[10] McDonald, D. "Steps Toward a Psycholinguistic Model of Language Production", MIT AI Lab Working Paper 193, MIT, Cambridge, Ma., April 1979.

[11] McKeown, K. R. "Paraphrasing Using Given and New Information in a Question-Answer System", In Proc. of the 17th Annual Meeting of the ACL. La Jolla, Ca., August, 1979, pp. 67-72.

[12] McKeown, K. R. "Generating Explanations and Descriptions: Applications to Questions about Database Structure," Technical Report # MS-CIS-80-9, University of Pennsylvania, Philadelphia, Pa., 1979.

[13] Simmons, R. and J. Slocum "Generating English Discourse from Semantic Networks." CACM. 15:10 (1972) pp. 891-905.

[14] Thompson, H. "Strategy and Tactics: A Model for Language Production" In Papers from the 13th Regional Meeting, Chicago Linguistic Society 1977.