COLING 82, J. Horecký (ed.)
North-Holland Publishing Company
© Academia, 1982

SURFACE ANALYSIS OF QUERIES DIRECTED TOWARD A DATABASE

Lawrence J. Mazlack
Richard A. Feinauer
University of Cincinnati
Cincinnati, Ohio, USA

A natural language interface is directed toward the database query languages that access machine stored data. A pattern driven transformation mechanism supports natural language access. A natural language is mapped onto a more formal computer database language. A human-like "understanding" of the query statement is not required. The transformation mechanism is separate from the target database management system. A goal is independence from both domain content and DBMS implementation. There is an emphasis on surface over content analysis.

Two particular questions are at issue. First, the extent to which a natural language interface to a database may operate independent of the subject domain of the database. Specifically, the extent to which natural language queries can be evaluated without the use of a query world descriptive reference system. Second, the extent to which natural language queries can be analyzed using pattern recognition techniques.

1 CONTEXT

As computer mass storage has become cheaper, more data has been stored in computers. The relationship between the stored data and access techniques has become increasingly more complex. This has resulted in the creation of very large databases and in the development of the powerful database management systems (DBMS) needed to access stored data. Most existing systems use a sophisticated data manipulation language (DML) to access the information in the database. These languages require detailed knowledge of the database's organization, the DML, and the host computer system to be used effectively. Computationally naive users must transmit their requests through a highly trained database expert who can use the DMLs. Simply put, data accessibility is limited by a communication barrier.

One way to have machine stored data directly avaliable to a wider range of people is to permit queries formulated in a natural language. Natural languages as data access languages have several compelling advantages. Some important ones include

1. A large number of potential computer users are unwilling or unable to learn and use formal machine languages.

2. For at least some applications natural language provides the ideal communications medium (Grishman and Hirshman,1978).

3. Potential users already know their natural language so little training in its use as a query language would be needed.

4. Natural languages are powerful tools for the expression of certain types of non-mathematical ideas and concepts.

5. The immediacy and flexibility of information retrieval are significantly improved when end users retrieve the data themselves.

If the user of computer stored data is able to access the data by using natural language, the utility of machine stored data would be increased. Not only would the casual user gain unhindered access, but an expert user could gain easier access as new DBMS statements would not have to be learned with every system change.

2 ACTIVITY OVERVIEW

Along with the general goal of developing a natural language database interface, a particular emphasis is on the portability of the interface. The goal is to achieve both domain and DBMS portability. By domain portability, we mean the capability to use the same natural language interface (NLI) to resolve queries against databases concerning different subject matter. By DBMS portability, we mean the capability to use the same NLI for a variety of DBMS implementations.

In order to achieve domain portability, it is clear that it is necessary to develop a system that minimizes and/or localizes the need for semantic referants. In order to achieve DBMS portability, it is necessary to limit contact between the DBMS mechanism and the NLI.

For the purposes of this paper, the term "syntax" will reference query surface structure and "semantic" will reference concerns which are not focused at surface structure. The focus of our semantic concerns are relatively narrow as the primary concern is with intentionality.

2.1 STRATEGY

General machine language processing has turned out to be difficult. It is unclear whether we currently have enough knowledge to develop a comprehensive machine natural language processing capability. Perhaps, the greatest opportunities for immediate success lie in the solution of subset problems. This investigation focuses on the relatively constrained natural language requirements necessary to support queries of a general database. Database queries require the capability to deal with a large subject context, but have a narrow pragmatic language use requirement. Others have sought to restrict problem complexity by trying to understand general statement about a limited world.

2.1.1 PHILOSOPHY

Our concern is with mapping of a natural language into a more formal language, not an understanding of the natural language. Hillman (1977) identifies thas as distinguishing between information retrieval and knowledge transfer. Knowledge transfer is dependent on access to knowledge representation systems capable of providing extensive help in gaining understanding. In comparison, in transformational mapping, if the question is HOW MANY DOGS ARE BLACK? a human-like understanding of the nature of DOGS is not at issue, but a way of formulating a database query to search the stored information with regard to the colour of dogs.

In focusing on the question of mapping from a natural language to a DML, the primary concern is not with the enhancement of understanding of language (as with Schank, 1973,1975) but rather attempting to bridge the gap between people and machines (as with Lehmann (1977), Ott(1977), Berrendonner(1980)).

2.1.2 PRAGMATIC AIDS

Natural language queries have three characteristics that aid their analysis: (1) the portion of a natural language´s syntax that must be covered is a subset of the entire language, (2) the pragmatic use of language in queries limits the interpretations of a statement, and (3) the analysis can be significantly guided by the assumption that a statement is a request for data from a known database.

2.1.3 PROBLEM REDUCTION

In separating the problems of access and DBMS design, both are simplified and made more amenable to solution. It would seem to be much more difficult to juggle with the problems of a natural language front end and at the same time to work on database development problems. The two problems would appear to compound each other.

Communication with a DBMS can be directed toward either the database structure or the actual contents of the database. In either case, communication flows through the DBMS facilities. We only consider content directed queries. To enhance the possibilities of DBMS portabilty, our NLI only makes contact with the data in the DBMS through the DML of the DBMS. Only the mapping between the final internal form of the NLI and the DML of the DBMS must be changed from DBMS to DBMS. The utilization of an existing database´s formal query language as the target representation allows the fundamental questions of query transformation to be addressed without the problems associated with with the collateral development of a DBMS.

2.2 OVERVIEW: QUERY TRANSFORMATION

Analysis of a query is treated as a transformation problem. The query is transformed from an informal language into a more formal language, the DML. The transformation is done in two steps: (1) the query is transformed from English into an internal representation and (2) the internal representation is transformed into a DML. The transformations are driven by a non-serial surface structure analysis. This analysis is supported by non-structural referants which are focused on recognizing the intended use of words and/or word groups.

The use of an internal intermediate representation of the query allows the determination of what is the desired information to be carried out in isolation from the peculiarities of the target DML. Also, by keeping the initial analysis of the query independent of the specific DML, the mechanism that puts the query in standard form will not have to be changed if the system is moved to a new DBMS with a different DML. This allows the analysis to be partitioned into two distinct phases: (a) transformation of the English query into a standard form and (b) subsequent construction of the DML query.

A simplified frames (Minsky,1975) type data structure called templates is used for use as the target internal representation. The analysis process includes identification of the template which best matches the query and the filling in of all the information needed to complete the stereotyped question.

2.3 OVERVIEW: THE ANALYZER

The analyzer in the mechanism uses both syntactic and semantic information to transform the query into an internal representation. Syntactic sources support as much of the analysis as is possible. When semantic sources must be used, existing sources of semantic information are used. This minimizes the amount of effort that must be expended in developing semantic referants. After the query is in a fully notated template representation, control of the mechanism passes to the bridge coding which transforms the standard form representation of the query into a DML.

2.4 THE BRIDGE CODE

The bridge coding transforms the query from a completed template into the DML of the host DBMS. Use of the DMLs of DBMSs has several advantages that are not exploited in systems which develope their own access routines. First, using the existing DMLs reduces the amount of new software that must be produced. Second, existing software such as report generators etc. would not require modification. Lastly, the DML can continue to be used directly, without going through the natural language processor, for those applications, such as updating, where the use of a natural language system may be undesireable.

After the template has been converted into a DML query, control of the system passes to the DBMS which will evaluate the query. When the DBMS is finished the answer and control of the system passes to a response generator.

3 THE ANALYZER

The analyzer transforms an English query into a semantically equilvalent template representation. The analyzer goes through four steps: a word role identifer, a phrase identifier, a phrase analyzer, and a template matcher. The template matcher is used to match template fragments to a template and to integrate the fragments into a single query. This approach is similar to the method used by (Wilks,1975a) in his preference semantics theory for general natural language processing.

Once the query is in a fully notated internal representation, the mechanism has established exactly what information the user requires. When this happens, the queries can then be transformed from the standard form into the DML of the host DBMS. To transform the query from English into an internal representation the analyzer has to identify in the query:

1. the desired information;
2. the required attributes;
3. any implied or assumed information.

Surface analysis of the query is used to do as much as is possible. From the analyzer, an understanding of the use of most of the words and word groups in the query is derived.

3.1 NON-SERIAL PROCESSING

Substantially all formal language analysis (compilers, etc.) procede serially (left-right or right-left). Also, most natural language parsing schemes procede on a serial basis. This is particularily true for natural language since Woods (1970) developed the powerful ATN concept. This project is significantly different in that it does not procede in a serial direction through a query. We resolve the easiest elements first, then use these resolutions to resolve the next element. By resolving an easy element first, where ever it is in the query, it often makes other query element resolution easier, irregardless of where the other element may be in serial relationship to the element resolved first. For example, if the query we are trying to resolve has the form

A B C D E

if the easiest element to resolve is C, it would be resolved first. Resolving C, might then make the resolution of B easy, etc.

3.2 IDENTIFING THE WORD ROLE

The problem of identifying the role of a word is not a trivial one since the same word may have a different role in different contexts. Some preliminary work on statistically- based identification (Mazlack,Feinauer,1980) has already been reported. Further to this, an identification mechanism using pattern recognition techniques has been developed. Initial word role labelling is supported by the use of various dictionaries and statistical data.

3.2.1 DICTIONARIES AND VOCABULARY

Several Dictionaries are applied successively. They are

(a) core dictionaries describing:
single role structural words (prepositions, conjunctions, question words, existance verbs, articals, quantifiers) and functional words (total, average, sum, etc.).
(b) terms appearing in the logical schema
(c) jargon
(d) a general dictionary

The dictionaries are used to provide canditate word roles. By applying these dictionaries, the words in a query are labelled. The process is reductive in that canditate word roles are reduced in number as successive dictionaries are applied.

A natural language interface which accepts a rich natural language input and reduces it to a constrained output, must reduce the variability of the words in the query. How the reduction is achieved is more than a simple table lookup with and attendent vocabulary reduction. A reduction down to a minimal set of words similar to Wilks (1975b) primatives is not required. Of more interest is the identification of a vocabulary which is not obviously redundant; i.e., with two words covering the same or nearly the same subject area. Vocabulary reduction takes place as part of both the word role identification and phrase recognition processes.

3.2.2 STATISTICS

A statistical knowledge about words in queries can contain such information as

(a) the chance that a word with a certain role will appear in a given position in a query

(b) the chance that a word with a certain role will appear in a specified position of a n-word pattern with the word roles in the other n-1 positions in the pattern specified

(c) the chance that a word with a certain role will appear after (or before) a specific vocabulary word

This information is used reductively to resolve words that had more than one canditate role after the dictionaries are applied.

3.3 PHRASES AND TEMPLATES

The method to be used to identify the phrase boundaries is a non-serial technique which uses keywords and the word roles identified by the word role identifier. This analysis is pattern driven and uses patterns developed from an extensive sample of actual DBMS queries.

Progressive recognition of the use of words and word groups leads to the development of patterns which include both syntactic and semantic groups. For example, the following patterns represent query template skeletons:

<ques. word><existance verb><desired info><verb>^{0}(<prep><reqd. attr.>)$^{+}$

WHAT<desired info><existance verb><reqd. attr.>(<prep><reqd. attribute>)$^{0,+}$

What happens is that the boundaries of a semantic group are delimited by syntactic units and a limited set of semantic purposes can be assigned to particular word groups in the input query. Once the pragmatic intentionality of a group is recognized, this group can then be further analyzed to identify the specific roles of words and word groups within it.

Once the initial analysis has been completed, phrase analysis mechanisms take over to transform the phrases into candidate template fragments. A template fragment contains that information in the phrase that is needed to accurately evaluate the query. This includes identification of what is the desired information, attributes the desired information must have and actions requested of the system. After the individual phrases have been transformed into template fragments the template matching mechanism takes over.

A pattern recognition approach selects the template that has the closest match between the information needed to complete the template and the information in the template fragments. A measure of fit "goodness" is developed and used to choose between competing interpretations. After the appropriate template has been selected the template matching mechanism completes the stereotyped query using information taken from the template fragments.

An example of this process can be found in the authors' paper: "A Pattern Driven Analysis of Queries Directed Toward Existing Databases" (Mazlack,Feinauer,1982).

3.4 SEMANTIC INFORMATION

The content of the database is not directly referenced as an information source. The logical database schema is used as a primary semantic information source because it already exists separate from the natural language query system and does not have to be created when the natural language analyzer is implemented with a new database. One cf the major problems problems with many existing natural language query systems is that they use significant information specific to the particular database they reference. By using information sources that do not have to be recreated for each new application, the amount of effort needed for new systems is reduced.

4.0 RECAPITULATION

The rapidly improving capabilities of computer hardware combined with the rapid decline in the cost of that hardware has created a situation where a major limiting factor in the utility and growth of DBMS is the inability of many people to use the complex software packages needed to access databases. One possible solution is the development of a natural language interface that can serve as an intermediary between the user and the DBMS. This would enable users to communicate with the database in their own language instead of the computers.

The mechanism described maximizes surface analysis of the query and minimizes the amount of content "understanding" needed to resolve a query. The construction of a world mechanism for each application Semantic information necessary to resolve a query is derived from existing sources such as the logical schema.

The research involved in completely specifying and implementing the mechanism is directed toward two fundamental questions. They are: (1) what is the minimum amount of "understanding" of a natural language query that is needed to generate a semantically equivalent DML query and (2) how much information about a query can be derived from surface analysis.

Note: A detailed comparison of this mechanism with other systems can be found in a working paper (Feinauer, 1981) on this mechanism.

5 BIBLIOGRAPHY

Barrendonner,A.,Bouche,R.,LeGuern,M.,Rouault,J. (1980) "Pour Une Methode D'Interaction Ponderee des Composats Morphologique et Syntaxique en Analyse Automatique du Francais," T.A. Informations, 1980, n1, p3-28

Codd,E.F. (1974) "Seven Steps To Rendezvous With The Casual User," DATA BASE MANAGEMENT, Klimbie,J.W., K.I. Koffeman (eds), North-Holland, 1974, Amsterdam, p179-200.

Feinauer,R.A. (1981)"A Proposed Natural Language Database Access Method," Working Paper, University of Cincinnati, 1981

Grishman,R., Hirschman,L. (1978) "Question Answering From Natural Language Medical Databases," Artificial Intelligence, 1978, v11, p25-43.

Hillman,D.J. (1977) "Model For The On-Line Management Of Knowledge Transfer," On-Line Review, v1, n1, March, 1977, p23-30

Lehmann,H. (1977) "The USL System for Data Analysis," PROCEEDINGS OF A WORKSHOP ON NATURAL LANGUAGE FOR INTERACTION WITH DATA BASES, Rahmstorf,G.,Ferguson,M. (eds), January, 1977

Mazlack,L.J., Feinauer,R.A. (1980) "Establishing A Basis For Mapping Natural Query Language," Proc. of The Joint British Computer Society and ACM Symposium: Research, Development In Information Retrieval, 1980, Cambridge, England.

Mazlack,L.J., Feinauer,R.A. (1982) "A Pattern Driven Analysis of Queries Directed Toward Existing Databases," European Conference on Artificial Intelligence, 1982

Minsky,M. (1975) "A Framework For Representing Knowledge," THE PSYCHOLOGY OF COMPUTER VISION, Winston,P. (ed), McGraw-Hill, 1975, New York, p211-278.

Ott,N.,Zoeppritz,M. (1977) "USL - An Experimental Information System Based on Natural Language," NATURAL LANGUAGE BASED COMPUTER SYSTEMS, Bolc,L. (ed), 1979

Plath,W.J. (1976) "Request: A Natural Language Question Answering System," IBM J. Research and Development, July, 1976, v20, n6, p326-335.

Schank,R. (1973) "Identification of Conceptualizations Underlying Natural Language," COMPUTER MODELS OF THOUGHT AND LANGUAGE, Schank,R., Colby,M. (eds), W.H.Freeman, 1973, San Francisco.

Schank,R. (1975) CONCEPTUAL INFORMATION PROCESSING, North-Holland, Amsterdam, American-Elseiver, 1975, New York, p1-82.

Wilks,Y. (1975a)"An Intelligent analyzer and Understander of English," CACM, May, 1975, v18, n5, p264-274.

Wilks,Y. (1975b) SEVEN THESES ON ARTIFICIAL INTELLIGENCE AND NATURAL LANGUAGE, ISSCO memo no. 17, 1975.

Winograd,T. (1973) "A Procedural Model of Language Understanding," COMPUTER MODELS OF THOUGHT AND LANGUAGE, Schank,R., Colby,M. (eds), W.H.Freeman, 1973, San Francisco, p152-186.

Woods,W.A. (1970) "Transition Network Grammars For Natural Language Analysis," CACM, October, 1970, v13, n10, p591-606.

Woods,W.A. (1977) "Lunar Rock In Natural English: Explorations In Natural Language," LINGUISTIC STRUCTURE PROCESSING, Zampolli,A. (ed), North-Holland, 1977, Amsterdam, p521-570.