ROBOT: A HIGH PERFORMANCE NATURAL LANGUAGE DATA BASE QUERY SYSTEM

Larry R. Harris
Dartmouth College
Mathematics Department
Hanover, New Hampshire 03755

Abstract

The field of natural language data base query has seen several successful research systems that have been able to process large subsets of the English language. The ROBOT system is a high performance natural language processor that extends the techniques developed in these earlier systems, in an attempt to provide a usable natural language query medium for non-technical users.

ROBOT is already installed in four real, world environments, working in five application areas. Analysis of log files indicate that between 80-90% of end user requests are successfully processed. The system successfully deals with both lexical and structural ambiguity, inter- and intra-sentential pronomilization and sentence fragments. The resource requirements of ROBOT are well within the limits of medium to large computer systems.

This paper describes the system from an information flow point of view. The goal is to obviate the creation of a semantic store purely for the natural language parser, since a different structure would be required for each domain of discourse. Instead the data base itself is used as the primary knowledge structure within the system. In this way the parser can be interfaced to other data bases with only minimal effort.

Sample Dialog

The following""" questions will give some idea of the current capabilities of the system. They refer to a personnel and car data base.

GIVE ME THE NAMES OF ALL EMPLOYEES WHO HAVE JOBS WORKING AS A SECRETARY IN THE CITY OF_ CHICAGO.

This question and the 3 that follow indicate the range of allowable expression, from verbose wordings to fragments.

PRINT THE NAMES OF ALL THE CHICAGO EMPLOYEES WORKING AS SECRETARIES.

WHO ARE THE CHICAGO SECRETARIES?

SECRETARIES IN CHICAGO.

The following sentences illustrate the lexical ambiguity of the word "and".

PRINT THE SALARY OF SMITH AND LAWLER.

WHO EARNS BETWEEN $20,000 AND $30,070?

PRINT THE NAME, PHONE AND FAMILY STATUS OF ALL THE NEW YORK STATE GIRLS.

FIND THE CARS MADE BY PORSCHE AND MADE IN '71.

This example illustrates how prepositions can be the critical factor in determining the field reference.

HOW MANY NH AND CT PEOPLE ARE THERE?

BROKEN DOWN BY STATE AND CITY, PRINT A SALARY REPORT FOR THESE PEOPLE, INCLUDING THEIR NAME AND AGE.

The following sequence illustrates how pronouns can be used to refer back to earlier points in the dialog.

WHO ARE THE SECRETARIES?

WHICH OF THEM LIVE IN CHICAGO?

There are now two valid pronoun referents.

WHICH OF THEM LIVE IN DETROIT?

In this case the pronoun referent is clearly the original set of secretaries, thus no clarification is required.

WHICH ONES ARE NOT MARRIED?

This question would require user clarification, since the user could be referring to all of the secretaries or just the Detroit secretaries.

Sources of Information

Fortunately, the contents of the data base define a sufficiently restrictive micro world such that only limited world knowledge is required by the parsing process. A very surprising and fortunate fact is that the data base itself can serve as this body of world knowledge. Furthermore, the existing structure of the data, in the comparatively simplistic form used by existing DBMSs, is adequate for data base query. In this section we discuss this entire question of how different kinds of information or world knowledge, are brought to bear on the understanding of a query. We consider there to be four distinct sources of information that can be applied to the understanding of a request. In order of their relative preference, we make use of:

1) the dictionary,
2) the sentence itself,
3) the data base, and
4) the user.

Let us discuss each of these in turn, and then discuss how they complement one another in providing enough information to understand the user's request.

The dictionary is the traditional table of word stems, syntactic categories, and semantic uses. Unfortunately it contains only localized information that must be pieced together to form a complete semantic structure.

The sentence itself is the primary source of information related to figuring out what the user requires. Initially most people believe that the sentence itself is all that is required to determine what the user wants. Unfortunately this is not the case, as many of the user's requests will be incomplete or potentially ambiguous. For example, consider the request, "TELL ME ABOUT THE FORDS'. This request is incomplete because it doesn't specifically tell what information the user wants about the "FORDS". It is ambiguous because the question may be interpreted in two totally distinct ways in the context of a file of presidents, or a file of cars. Information

from outside the sentence must be brought to bear in these situations, to fill in the unspecified portions of the request and to disambiguate the various interpretations.

Putting aside these degenerate cases of requests that are virtually information free, the sentence is in fact, typically a very rich source of information. The information in the sentence is not quite so accessible as that in the dictionary. The distinction is that in order to extract information from the sentence, we have to parse it first.

The use of the data base itself as a knowledge structure is a major reason why the ROBOT system has been able to* achieve its level of performance. The data base is used to provide information of two sorts. First, it can tell exactly how individual words are used in the data base. Second, it can tell which interpretations of a given request make sense in the current state of the data base. In Harris[77b] we discussed these issues in detail.

The most beneficial aspect of the use of the data base as a source of information is that it provides the perfect complement to the information contained within the sentence. That is to say, that it is in just those cases where the information in the sentence is incomplete or ambiguous, that the data base provides just the right answers. An example will illustrate this point. "WHO ARE THE NEW YORK EMPLOYEES LIVING IN BUFFALO?" The parse of this sentence yields little information about the relationship between "NEW YORK" and "EMPLOYEES" because this is noun-noun modification, which could allow almost any type of semantic relationship. It is here that the data base helps out by informing us that "NEW YORK" is a data item in the city and state fields. This is enough to allow us to build the necessary semantic structure for the request.

But the problems with understanding the request are not yet over because of the ambiguity of "NEW YORK". There is not enough information within the sentence itself to disambiguate the meaning of "NEW YORK". Once again, the data base is instrumental in determining that the logical intersection of "NEW YORK STATE" and "BUFFALO" makes sense within the data base, whereas "NEW YORK CITY" and "BUFFALO" does not. With this additional piece of "world knowledge" we can generate a unique semantic interpretation of the request. It should be clear from the foregoing discussion that if the original request had been more information rich, such as "WHO ARE THE NEW YORK STATE EMPLOYEES LIVING IN BUFFALO", then this last interaction with the data base would have been avoided.

The user himself is our last resort in terms of obtaining information in order to understand the request. We put the user last on the list in order to avoid bothering him unless absolutely necessary, and to do as much work for him as we possibly can. Had we gone to the user and not the data base in the last example, we would have required the user himself impart the fact that Buffalo is in New York, when that piece of information is derivable from the data base. We prefer to apply all the information at our disposal before bothering the user.

There are, of course, certain times when the user's request is so ambiguous that we must ask him what he meant. For example the request, "WHAT IS GREEN?" would require clarification from the user in a data base environment vhere "GREEN" is both a name

All four sources of information are brought to bear on the problem in the following way. First, the dictionary is used to recognize individual words. In some cases an individual word may consist of several actual words, such as "NEW YORK". This determination is made wherever possible by interrogating the dictionary. In other cases the data base must be consulted to recognize word pairings that occur in the data base. At this point the sentence is parsed, yielding in general, several incomplete interpretations. The data base is used to fill in the incomplete portions of the interpretations, and to eliminate those interpretations that don't make sense in the current state of the data base. If more than one interpretation still persists, then the user is consulted,

Summary

The fact that existing data base structures are capable of resolving most of the semantic issures of natural language query has allowed for the existing AI parsing techniques to be feasibly applied in real world applications. As an example of what can be done with existing natural language processing techniques, the ROBOT system serves as a counterexample to the philosophy that dealing with natural language is an all or nothing proposition. As an example of what can be done within the existing data base technology, the ROBOT system defines a baseline of performance, against which future systems embodying a more powerful semantic structure, can be measured.

References

Harris[72] L.R., "A Model for Adative Problem Solving Applied to Natural Language Acquisition", TR 72-133, Department of Computer Science, Cornell University.

Harris[77a] L.R., "ROBOT: A High Performance Natural Language Interface for Data Base Query", TR 77-1, Mathematics Department, Dartmouth College

Harris[77b] L.R., "Natural Language Data Base Query: Using the data base itself as a definition of world knowledge and as an extension of the dictionary", TR 77-2, Mathematics Department, Dartmouth College