THERE STILL IS GOLD IN THE DATABASE MINE

Madeleine Bates

BBN Laboratories
10 Moulton Street
Cambridge, MA 02238

Let me state clearly at the outset that I disagree with the premise that the problem of interfacing to database systems has outlived its usefulness as a productive environment for NL research. But I can take this stand strongly only by being very liberal in defining both "natural language interface" and "database systems".

Instead of assuming that the problem is one of using typed English to access and/or update a file or files in a single database system, let us define a spectrum of potential natural language interfaces (limiting that phrase, for the moment, to mean typed English sentences) to various kinds of information systems. At one end of this spectrum is simple, single database query, in which the translation from NL to the db system is quite direct. This problem has been addressed by serious researchers for several years, and, if one is to measure productivity in terms of volume, has proved its worth by the number of papers published and panels held on the subject. Indeed, it has been so deeply mined that the thought "Oh, no! Not another panel on natural language interfaces to databases!" has resulted in this panel, which is supposed to debate the necessity of continuing work in this area rather than to debate technical issues in the area. And yet if this problem has been solved, where is the solution? Where are the applications of this research?

True, commercial natural language access interfaces for some database systems have been available for several years, and new ones are being advertised every month. Yet these systems are, now, not very capable. For example, one of these systems carried on the following sequence of exchanges with me:

User: Are all the vice presidents male?
System: Yes.
User: Are any of the vice presidents
      female?
System: Yes.
User: Are any of the male vice presidents
      female?
System: Yes.

Nothing was unusual about either this database or the corporate officers represented in it. The system merely made no distinction between "all" and "any", and interpreted the final query to mean the same as "Are there any vice presidents who are either male or female". This same system, when asked for all the Michigan doctors and Pennsylvania dentists, produced a list of all the people who were either doctors or dentists and who lived in either Michigan or Pennsylvania. This is the state of our art?

But, you are probably thinking, those examples don't illustrate research problems that need to be worked on; they are problems that were "solved" years ago. But I contend that it is not enough to strip broad areas of research and develop isolated theories to account for those areas, because the result is similar to that of strip mining coal: local profit followed by more global losses. It is more beneficial to choose a limited area (such as database interfaces, perhaps extended a bit as described below) and mine it very deeply, not necessarily discovering every aspect of the domain but requiring that the various aspects be integrated with one another to produce a coherent whole.

Even in the most simple database access environment, one can find in natural queries and commands examples involving meta-knowledge ("What can you tell me about X?"), presupposition (Q: "How many students failed Math 108 last semester?" A: "Math 108 wasn't given last semester."), and other not-yet-mined-out topics. Extending the notion of database access to one of knowledge-base access where information may be manipulated in more complex ways, it is easy to generate natural examples of counterfactual conditionals ("If I hadn't sold my IBM stock and had invested my savings in that health spa for cats, what would my net worth be now?"), word sense ambiguity (the word "yield" is ambiguous if there is both financial and productivity data in the knowledge base), and other complex linguistic phenomena.

Let us go on to define the other end of the spectrum I began to explicate above. At this end lies a conversational system for query, display, update, and interaction in which the system acts like a helpful, intelligent, knowledgeable assistant. In this situation, the user carries on a dialogue (perhaps using speech) using language in exactly the same way s/he would interact with a human assistant. The system being interfaced to would, in this case, be much more complex than a single database; it might include a number of different types of databases, an "expert system" or two, fancy display capabilities, and other goodies. In this environment, the user will quite naturally employ a wider variety of linguistic forms and speech acts than when interfacing to a simple db system.

One criticism of the simple db interfaces is that the interpretive process of mapping from language concepts onto database concepts is sufficiently unlike the interpretation procedures for other uses of natural language that the db domain is an inappropriate model for study. But not all of the db interfaces, simple or more complex, perform such a direct translation. There is a strong argument to be made for understanding language in a fairly uniform way, with little or no influence from the fact that the activity to be performed after understanding is db access as opposed to some other kind of activity.

The point of the spectrum is that there is a continuum from "database" to "knowledge base", and that the supposed limitations of one arise from the application of techniques that are not powerful enough to generalize to the other. The fault lies in the inadequate theories, not in the problem environment, and radically changing the problem environment will not guarantee the development of better theories. By relaxing one constraint at a time (in the direction of access to update, one database system to many, a database system to a knowledge-based system, simple presentation of answers to more complex resonses, static databases to dynamic ones, etc.), the research environment can be enriched while still providing both a base to build on and a way to evaluate results based on what has been done before.

Some Research Issues Related to Databases

Here are a few of the areas which can be considered extensions of the current interest in database interfaces and in which considerable research is needed. Large, shiny nuggets of theory are waiting to be discovered by enterprising computational linguists!

1. Speech input. Interest in speech input to systems is undergoing a revival in both research and applications. Several "voice typewriters" are likely to be marketed soon, and will probably have less capability than the typed natural language interfaces have today. But, technical and theoretical problems of speech recognition aside, natural spoken language is different linguistically from natural written language, and there remains a lot of work to be done to understand the exact nature of these differences and to develop ways to handle them.

2. "Real language". By which is meant (written or spoken) language complete with errors, ungrammaticalities, jargon, abbreviations, telegraphic compression, etc. Research in these areas has been going on for some time and shows no sign of running dry.

3. Generating language. An intelligent database interface assistant should be able to interject comments as appropriate, in addition to displaying retrieved data.

4. Extended dialogues. What do we really know about handling more than a few sentences of context? How can a natural conversation be carried on when only one of the conversants produces language? If able to generate language as well as to understand it, a database assistant could carry on a natural conversation with the user.

5. Different types of data bases and data. By extending the notion of a static, probably relational, database to one that changes in real time, contains large amounts of textual data, or is more of a knowledge base than a data base, one can manipulate the kind of language that a user would "naturally" use to access such a system. for example, complex tense, time, and modality expressions are almost entirely absent from simple database query, but this need not be the case.

All of this is not to say that all the research problems in computational linguistics can be carried on even in the extended context of database access. It is rather a plea for careful individual evaluation of problems, with a bias toward building on work that has already been done.

This environment is a rich one. We can choose to strip it carelessly of the easy-to-gather nuggets near the surface and then go on to another environment, or we can choose to mine it as deeply as we can for as long as it is productive. Which will our future colleagues thank us for?