

PANEL

NATURAL LANGUAGE AND DATABASES, AGAIN

Karen Sparck Jones

Computer Laboratory, University of Cambridge
Corn Exchange Street, Cambridge CB2 3QG, England

INTRODUCTION

Natural Language and Databases has been a common panel topic for some years, partly because it has been an active area of work, but more importantly, because it has been widely assumed that database access is a good test environment for language research. I thought the time had come to look again at this assumption, and that it would be useful, for COLING 84, to do this. I therefore invited the members of the Panel to

 speak to the proposition (developed below) that database query is no longer a good, let alone the best, test environment for language processing research, because it is insufficiently demanding in its linguistic aspects and too idiosyncratically demanding in its non-linguistic ones;

and to

 propose better task environments for language understanding research, without the disadvantages of database query, but with its crucial advantage of an independent evaluation test.

DATABASES: PROS, CONS, AND WHAT INSTEAD?

Database query has a long and honourable history as a vehicle for natural language research. Its value for this purpose was restated, for example, by Bonnie Webber at IJCAI-83 (Webber 1983). I nevertheless think it is now time to question the value of database query as a continuing vehicle for language research.

Database query has two major points in its favour. The task is relatively restricted, so success in building a front end does not depend on solving all the problems of language and knowledge processing at once. More importantly, the task provides a hard, rather than soft, test environment for a language processor: the processor's performance is independently evaluated via its output formal search query.

Natural language research has profited in the past from the restrictions on the database task: its limited linguistic functions and world references have allowed concentration on, and hence progress in dealing with, obvious problems of language and knowledge processing. But I believe that database query is reaching the end of its utility for fundamental research on natural

language understanding, for two reasons.

The first is that current database systems are too impoverished to call for some important language-processing capabilities in their front ends, so work on these capabilities is discouraged. Obvious examples of the expressive poverty of typical database systems include their lack of resources for handling, at all properly, such important components of text meaning as qualifying concepts like negation and a variety of quantifiers; intensional concepts including meta description, modality, presupposition, different semantic relations, and constraints of all sorts; and the full range of linguistic functions subsumable under the heading of speech acts. More generally, the nature of the task means that many typical requirements of language understanding, e.g. the determination of the domain of discourse and hence senses of words, and many typical forms of language use, e.g. interactive dialogue, are never investigated. (Though attempts may be made, forced by the way natural language is actually used in input, to handle some of these phenomena via superimposed knowledge bases, this does not undermine my general point: the additional resources are merely devices for reducing the richness of natural language expressions to obtain sensible database mappings.)

The second reason for doubting the continuing utility of database query as a field for natural language research, is that the autonomous characteristics of database systems impose idiosyncratic constraints on the language processor that are of no wider interest for natural language understanding in general. Most of the problems listed by Robert Moore at ACL-82 (Moore 1982) fall into this class, as do many of those identified by, for example, Templeton and Burger (1983). The examples include database-specific quantifier interpretation, quantity determination, procedures for mapping to compound attributes, techniques for dealing with open value word sets, and ripping apart complex queries. Further, even more database oriented, problems include, for instance, path optimisation, parallel (coroutine based) query evaluation, and null values.

These problems can be very intractable for individual data models or databases, and as the solutions tend to be ad hoc and specialised, the issues are essentially diversions from research on more pervasive language phenomena and functions, and hence on generally relevant language understanding procedures.

This is of course not to deny that database access presents many perfectly 'ordinary' language interpretation problems. The crux is whether the central interpretive process, mapping from language concepts onto database ones, is sufficiently like the interpretation procedures required for other natural language using functions, for it to be an appropriate study model for these.

I believe that much of the attraction of the database case comes from the stimulus to logic-based meaning representation provided by the formal database query languages into which natural language questions are usually ultimately mapped. The database application naturally appeals to those who believe that the meanings of natural language texts should be expressed in something like first order logic.

But current data languages, however logical, are very limited. More importantly, they are geared to data models expressing properties of databases that are manifestly artificial, and are not properties of the real worlds with which natural language is concerned. Third normal form is a property of this kind. I do not believe that third normal form has got anything to do with the meaning of natural language expressions. But the ultimate consequence of working with present data models is behaving as if it does. This is clearly unsatisfactory. I am of course not attacking the idea of logical meaning representations. What I am claiming is that the database application is an inadequate test environment for natural language understanding systems.

One argument for continuing with database query processing must therefore be that those mainstream language handling problems which do arise have not been fully resolved, so it is legitimate to concentrate on these, in what is a convenient test environment, and defer an attack on other language processing tasks. The second is that there are ill-understood knowledge handling operations triggered by and interacting with language processing that are not specialised to one contemporary computational task, but are sufficiently typical of a whole range of other knowledge processing tasks to justify further study in the exemplary database case.

Without wishing to imply that the database query function is all wrapped up (or doubting the need for much further system engineering), I do not think these arguments are strong, simply because it is impossible to disentangle general language problems from database ones, and database problems from current highly restricted data models and implementations. Moore's example of time and tense illustrates this very well. Time information determination problems arise in database questions; but because of the database domain context, they are typically only an arbitrary subset of those ordinarily occurring, and require interpretive responses biased to the particular time concepts of the database. It may be that finding anything out about time interpretation, even in a limited context, is of some use. But it is surely better to consider time interpretation in the more motivated way allowed by a richer environment involving a fuller range, or at least less arbitrarily selected set, of temporal concepts than

those of current databases.

My point is that to make progress in natural language research in the next five to ten years we need the stimulus of a new application context. This must meet the following criteria: it must be more 'central' to language understanding than database query; it must be harder, without overwhelming us with its difficulty; and we should preferably be able to make a start on it by exploiting what we have learnt from the database application. But most importantly, the new task must have built-in evaluation criteria for the performance of language processors. This is more difficult to achieve with systems whose entire function is language processing, like translation, than with systems where natural language processing is required for the system's external world interface; but it is still possible to evaluate translation, for example, or summarising, reasonably objectively: the problem is the sheer effort involved.

Some candidate applications meeting these criteria are:

- natural language interfaces to conventional computing systems (e.g. operating systems, numerical packages, etc.)
- natural language interfaces to expert systems
- natural language interfaces to robots
- natural language interfaces to teaching systems

All of these meet the evaluation requirement; what requires examination is the extent to which non-trivial back end systems (e.g. a robot more interesting than SHRDLU) would be too severe a challenge for language processing. It is not necessary, in this context of principle, to base choices on potential market interest: expert systems would score here, presumably. However it is necessary to consider the expected 'technological' plausibility for the requirement for a natural language interface e.g. to a robot.

These candidates are for interface systems. Should we instead be renewing the attack on language systems, e.g. for translation or summarising; or upgrading semi-linguistic systems like those for document retrieval?

REFERENCES

- Webber, B.L. 'Pragmatics and database question answering', IJCAI-83, Proceedings of the Eighth International Joint Conference on Artificial Intelligence, 1983, 204-205.
- Moore, R.C. 'Natural-language access to databases - theoretical/technical issues', Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics, 1982, 44-45.
- Templeton, M. and Burger, J. 'Problems in natural-language interface to DBMS with examples from EUFID', Proceedings of the Conference on Applied Natural Language Processing, 1983, 3-16.