EVALUATION OF NATURAL LANGUAGE INTERFACES TO DATABASE SYSTEMS: A PANEL DISCUSSION

Norman K. Sondheimer, Chair
Sperry Univac
Blue Bell, PA

For a natural language access to database system to be practical it must achieve a good match between the capabilities of the user and the requirements of the task. The user brings his own natural language and his own style of interaction to the system. The task brings the questions that must be answered and the database domain's semantics. All natural language access systems achieve some degree of success. But to make progress as a field, we need to be able to evaluate the degree of this success.

For too long, the best we have managed has been to produce a list of typical questions or linguistic phenomena that a system correctly processed. Missing has been a discussion of their importance and a similar list of unhandled phenomena. Only occasionally were even informal evaluations of systems conducted.

Recently, this has begun to change. In the last several years, many of the current generation of natural language access to database systems have been subject to laboratory or field testing. These include INTELLECT, LADDER, PLANES, REL and TQA. We have begun to discover what a user will ask a system, how he reacts to its limits, and where we need further work.

This panel brings together a good sampling of the people involved in these tests including individuals intimately involved with the above systems. The position papers that follow present their unique viewpoints on the important issues in the evaluation of natural language access to database systems. These include:

I. What has been learned about a) user needs, b) system's capabilities and c) their match with respect to tasks. Under this, what are the most important linguistic phenomena to allow for? What other kinds of interactions, beside retrievals, do users request? How good are systems at satisfying users? How good are users at finding ways to use systems? How satisfied are users with systems' performance? How does these results vary with respect to tasks?

II. What have we learned about running evaluations? Under this, what methodologies are capable of revealing what sorts of facts? What are the limits of field studies versus controlled experiments? How good are studies with a simulated system, such as Malhotra's with its human intermediary[1]? What are the independent variables that must be allowed for? What tools are available to determine user bias and experience beforehand, and user satisfaction afterward?

III. On the basis of these evaluations, what should the future look like for natural language access to database? Under this point, what niches look most promising for natural language interfaces? What standards should be set for natural language systems performance? What kinds of evaluations should be run in the future? How should they be designed and how should they be judged?

In addition to the position papers that follow, I strongly urge you to consult the panelist more extensive publications.

Bibliography

[1] Malhotra, A., "Design Criteria for a Knowledge-Based English Language System for Management: An Experimental Analysis", Ph.D. Thesis, MIT, MAC TR-146, 1975.