TWO SEMANTIC WORLDS: A DATA BASE SYSTEM WITH PROVISION FOR NATURAL LANGUAGE INPUT

P. Del 1'Oreo
IBM Italia - Bar! Scientific Center - Via Cardassi, 3 - 70122 Bari ITALY

Bar! University - Informatics Departement - Palazzo Ateneo - 70100 Bari ITALY

M. King
ISSCo - Institute for Semantic and Cognitive Studies - 17 Rue de Candolle
1205 Geneve SWITZERLAND

V.N. Spadavecchia
IBM Italia - Bari Scientific Center - Via Cardassi,3 - 70122 Bar! ITALY

This note very briefly describes a data base system with provision for Natural Language Input. No attempt is made here to justify the strategies adopted. Those interested In justification are referred to (Dell'0rco/ King, Spadavecchia 1977).

The system has been deliberately designed to be very modular. The data base itself Is a relational model based on Codd (Codd 1970).Two sample data bases are considered in the present system prototype: the first is about energetic resources and the second is about a department store. This last contains information about items sold and supplied, employees, departments,etc...(detai 1 s in (Chamberlin 197 4)). The data base Is interrogated by means of a formal query language (AQL) which is interpreted into a set of APL procedures. A user experienced in computer usage would be able to use AQL directly if he chose, thus avoiding natural language input and economising on processing time. AQL and the data base Implementation form two of the main modules of the system. Details of AQL, of its implementation and of that of the data base management system can be found in (Antonacci,Del 1 'Oreo, Spadavecchia 1976).

Obviously the most interesting feature of the system from an AI view point is the natural language input. A critical theorical decision has been made here. It Is by now generally agreed that any adeguate natural language processing system must make use of general world knowledge. In a data base system the world knowledge involved seems to us to break down Into two parts: the general knowledge of the world encapusulated in language use, and the much more restricted knowledge of the data base world. Thus words and phrases quite legitemately used in the natural language formulation of a question In a certain sense may have no meaning within the data base world. Before such a meaning can be attributed to them a correspondence must be set up between them and the formal objects of the data base.

In order to accomplish this two semantIcally driven modules deal with natural language analysis. The first is a general natural language analyser intended to allow as wide a subset of natural language as possible. It Is Independent of any particular data base, being based on the semantic structures of the natural language. This module is a development of a part of Wilks* Preference Semantics System (Wilks 1975), and establishes an intermediate semantic representation very similar to that used by Wilks as Input to the generation section of his system (see Herskowitz 1973). However, some difficulties of reference and of disambiguation which, for Wilks, would be resolved by common sense inference rules, remain in our representation to be dealt with by the next module. This intermediate representation serves as Input to the second semantIcally driven component. But now the semantics involved is the semantics of the data base, not the semantics of the natural language.

Connections between the elements of the intermediate representation and the names of relations and domains of the data base *are* established. For example, In a question about departments which sell shoes, the word "shoe" has attached an Information specifying that it is an item which can be sold or supplied. The verb "to sell", connected to "shoe", helps to choose the right data base *reference* for "shoe". A Transformation algorithm then recodes the question In terms of the formal query language mentioned earlier. Thus, when the data base is changed, unless the vocabulary dealt with by the first semantic module needs to be enlarged, all that needs to be changed are the connections between the elements of the natural language and the data base.

The analysis algorithm and the encoding algorithm are totally unaffected by a change in the data base. Similarly an expansion of the data base or a change in its structure do not affect the basic analysis algorithms.

This seems to offer great advantages, since the major effort in developing such a system inevitably goes into the 'permanent[1] parts-the natural language analyser, the encoder, the design and implementation of the formal query language and efficient storage, interrogation and manipulation of the data base itself.The ultimate goal is to have a system which can be used by the non DP specialist to access data using his natural way (in our case, the language) to describe the characteristics of the data to be retrieved. At the present stage of this work the entire formal query language has been implemented, while the Natural Language analyzer has been fully designed and Is under development.

(1) Antonacci F., P. Dell'Orco, V.N. Spadavecchia - AQL: an APL Based System for Accessing and Manipulating Data in a Relational Data Base System - Proceedings APL '76 Conference, ACM, New York, 1976, 31-42.

(2) Chamberlin D.D., R.F. Boyce SEQUEL: a Structured Englsh Query Language - IBM Research Report, RJ 1394, May 24,1974.

(3) Codd E.F. - A Relational Model of Data for Large Shared Data Banks - CACM 13. 6, June 1970, 377-387.

(U) Dell'Orco P., M.King, V.N. Spadavecchia - Catering for the Experienced and the Naive User: a Data Base System with Natural Language Query Facilities Proceedings of the IIASA Workshop on Natural Languages for Interaction with Data Base. (in print). IIASA - Jan.10-14, 1977 - 2361 Laxenburg - AUSTRIA

(5) Herskowitz A. - The Generation of French from a Semantic Representation, A.I. Laboratory Memorandum (1973), Stanford University, Stanford, California.

(6) Wilks Y. - An Intelligent Analyzer and Understander of English - CACM I8, 5, May 1975, 264-274.