SAPHIR+RESEDA, A HEW APPROACH TO INTELLIGENT DATA BASE ACCESS

Bernard Euzenat
Bernard Normier
Antoine Ogonowikl
ERLI, 72 qual des Carrieres. 94220 Charenton. France

Glan Ptero 2arr1
Centre National da la Racharchs Scientifique. Paris, Franca
and
TECSIEL. via Barnaba Oriani 32. 00197 Roma. Italy

ABSTRACT

This paper describes a transportable natural language interface to databases, augmented with a knowledge base and inference techniques. The inference mechanism, based on a classical expert system's type of approach, allows, when needed, to automatically convert an Input query into another one which 1s "semantically close". According to RESEDA'S theory, "semantically close" means that the answer to the transformed query Implies what could have been the answer to the original question. The presented system Integrates natural language processing, expert system and knowledge representation technology to provide a cooperative database access.

I. INTRODUCTION

Most of the existing natural language interfaces (NLIs) to databases - too numerous to be cited - free the user from learning the cumbersome and difficult syntax of formal database query languages. Still they impose a very Strict semantic model, which corresponds to the way a real world domain Is represented 1n a particular database (DB).

It is therefore difficult to produce a "valid question" without being aware of the structure of the database concerned. This issue is the main concern of the SAPHIR+RESEDA (Euzenat et al.. 1984a. 1984b) research project at ERLI.* SAPHIR (Normier, 1984, Normier et al.. 1984) our transportable, domain independent NLI to databases, is in the course of being provided with a relatively general knowledge base (KB) and associated inference mechanisms.

Before going into the technical details of this new system, let us show a typical problem that SAPHIR+RESEDA will be able to handle.

Suppose that you want to ask a question such as "Who has ever been in the US", and that youn staff management database does not contain any Information concerning the employees' excursions. Suppose further that the DB knows instead where each one of them was born and what their degrees are,

Using simple common sense knowledge, without accessing the database, we can tell the user that although we do not explicitly know who has been in the USA, we can nevertheless look for people born there and/or having an American degree. If the user accepts this proposal then we transmit the modified query to the DBMS

• This research was supported in part by the "Agence de Informatique" research contract ADI-84-068

Figure I contains a simplified flowchart of our system capable of such rational behaviour

```
+--------------------------------------------------------*
|   natural language query                               |
|              |                                         |
| |          GENERAL      PARSER              |          |
| |(domain independent but language dependent )|         |
|              |                                         |
|  deep predicate representation                         |
|+------>------»                                         |
||_____________|_________________________________________|
||  / does this have  °                                  |
||  /a representation "___ yes___Query the DB            |
|| ° in the particular /                                 |
||  D   database ,?    /                                 |
||   c______________/                                    |
||             | no                                      |
||   |    TRANSFORMATION MECHANISM   |                   |
|||          in the "RESEDA style"   |                   |
||   | (uses a "general" knowledge    base)|             \
||             |                                         |
||  a "different" deep predicate representation          |
|       |      |                                         !
```

Figure 1

First of all the natural language (\ e French query is parsed and partially disambiguated, using structure rules (Gross. 1975) and a syntactical grammar of French In some cases the parser resorts to simple clarifying dialogues As the result of this analysis, a predicate representation is produced It carries the meaning of the user's query Due to space limitations the parser will not be presented here

The system has at its disposal a model of tne structure of the database being accessed (a special predicate calculus description, close to the conceptual schema) so it can check whether this Query is a meaningful request to the database If it is meaningful, then the syster translates the deep predicate representation into a corresponding formal query language formula (this is how our existing NLI - SAPHIR proceeds, currently we can produce QBE SQL, AdaDas and Clio translations)

However, in many casts the query will not map directly onto the database.

Instead of simply telling the user that his question 1s not "the question to ask", we will try to transform his request Into something meaningful In the particular database and semantically close to the original request - this part of the system is based on G.P. Zarri's RESEDA system (Zarn. 1983. 1984a. 1984b. 1984c) We define "question1" to be semantically close to "questlon2" Iff the answer to "questlon1" Is Implied by the answer to "question2".

III. THE KNOWLEDGE BASE

To achieve this, we need of course the knowledge of what is "semantically close".

In the example presented above the system was able to transform the original question, because it knew that "If a person Is born In a particular place then that person must have been In that place at least *for* a short period of time".

The second alternative of the transformed query is produced because the system knows that "in order to obtain a degree, a person has to attend courses at a particular university or do some research work at that university" and that "in both cases the person had to be physically present there at one time or another" (to simplify the presented example we deliberatly ignore all the cases of exception to this rule).

It follows from the example that the system manipulates rather general common sense knowledge.

We decided to take the domain of staff management as our test-bed. However we still need to represent some more general, extra-domain knowledge

From a technical point of view, our overall knowledge base Is divided Into a number of smaller knowledge bases. as illustrated In Figure 2.

```
+--------------------------------------------------+
|____________________________                      |
|| Knowledge Base            |                     |
||  ________________________ | ex  a human being   |
||  |KB1 (most general)    | |     has a sex       |
||  ------------------------ |                     |
||  ________________________ | ex  the staff       |
||  |KB2 (domain specific) | |     management      |
||  ------------------------ |     legislation     |
||  ________________________ |                     |
||  |KB3                   | | ex  particular rules|
||  |(application specific)| |     of staff manage-|
||  ------------------------ |     ment for a parti-|
||___________________________|     cular company   |
+--------------------------------------------------+
```

Figure 2

KB1 contains bits of knowledge that are usable in any application (of course. It will *never* be complete)

KB2 contains general knowledge for a particular domain We hope to develop a certain number of KB2s (with a common KB1 nucleus) for each real world domain that our system will nave to deal with

Each KB1 ♦ KB2 couple is transportable "as 1s". as long as we do not switch to a completely different domain

Note that KB3 is not specific of a particular database, the system can use the same KB3 1n order to access different personnel databases of the same company.

The model of the database itself (the most specific level of knowledge) Is not included here since 1t does not take part directly In the transformation process.

IV. TYPES OF KNOWLEDGE

Any one of the KBs mentioned above may contain two types of knowledge: factual knowledge and rules (represented In a formalism based on RESEDA).

A. Factual knowledge

A fact such that "a university is in a city...", which Is a piece of factual knowledge. 1s represented by the highlighted part of Figure 3.

```
+--------------------------------------------------------+
| ...                                  ...               |
|  ^                                    ^                |
|  |                                    |                |
| isa                                  isa               |
|  |                                    |                |
|SCHOOL       GEOGRAPHIC_AND_ADMINISTRATION_UNIT         |
|  ^                  ^                  ^               |
|  |                  |                  |               |
| isa                isa                isa              |
|  |                  |                  |               |
|UNIVERSITY--loc. in--> CITY --loc. in--> COUNTRY        |
|   \--- ...--->                                         |
+--------------------------------------------------------+
```

Figure 3

This frame-like network describes "specific-generic" type of relations ("isa" links, we allow polyhierarchies) used as matching criteria when applying rules (cf below) Other specific relations (used in the rules, for example topological ones) are also present in the network.

This representation does not have the full power of frame systems since presently there is no procedural knowledge represented in the network

B. Rules

Currently we have two types of rules:

1 Transformation rules
2 Standardization rules

A fact such as "if a person (x) has received ("BE-AFFECTED-BY") a degree in a particular place (y) then that person has been ("BE-PRESENT") in that place", which is a transformation rule, is represented as in Figure 4

```
+--------------------------------------------------+
|        BE-PRESENT             BE-AFFECTED-BY     |
|           SUBJ x   y   -->      SUBJ x           |
|                                  OBJ  degree   y  |
|                                                  |
|        y  is an indication of location           |
+--------------------------------------------------+
```

Figure 4

Note that:

a. This is a deep case representation with a limited number of predicates ("deep verbs").

b. The direction of the arrow is following RESEDA'S convention. From a practical point of view the arrow indicates that the right-hand side schema is allowed to replace something that matches the left-hand side schema. Conceptually 1t means that the Information possibly retrieved using the right-hand side schema Implies the Information searched for. by the (original) left hand side schema.

Rules can place restrictions on variables (as Is the case in RESEDA). For Instance, the rule shown above could have required that "x" be a human being. This rule would then apply for "x ■ y". where "y" is known to be a teacher. The match (between x and y) 1s made possible, because we have stored 1n the factual knowledge representation the fact that "teacher '1s_a' human_be1ng".

Standardization rules are used to translate the parser's output (containing surface verbs) into a more canonical form. i.e. usable by transformation rules. All of the standardization rules belong to the KB1 level, as they are absolutely general.

V. THE INFERENCE ENGINE

We are In the course of developping (December 84) a prototype version with an "exhaustive" engine. meaning that all the possible transformations are executed (with simple destructive chronological backtracking) until a representation is found. This approach is very close to the RESEDA Inference engine (see Zarrl. 1984c. for a description of the later one)

This simplistic approach 1s justified since our current knowledge base contains only "KB3" types of knowledge.** We are planning to move onto more sophisticated approaches (ex. choosing among several representations in the database of the same question, a.so...) after the prototype has been well tested.

VI. CONCLUSIONS

The presented NLI system is designed to tackle issues such as transportability and "helpful understanding" of a naive user

Transportability, the first Issue. has been examined in many research systems. however all of these seem to lack a clear distinction between different levels of transportability of the knowledge represented (HAM-ANS. Hoeppner et al. . 1983 seems to be an exception)

In addition SAPHIR+RESEDA 1s using general common sense knowledge (reference Hemphill and Rhyne. 1978. describes a project where the use of Schank's general formalism 1n DB queries has been explored, but its goals seem to be more limited than ours).

The second issue is somewhat close to " cooperation" as defined in the COOP system (Kaplan. 1982) Kaplan however seems to consider only the case where the DB produces a null answer. We believe his and our approach to be complementary.

Our system uses rules to produce a DB query, this 1s similar to a deductive DBMS approach. However our rules are not formal and we *never* access the database in the course of transformation (because In most real world cases 1t 1s prohibitively expensive).

** Our rules (temporarily) have a very specific content. I.e. they have not been "factored out". This 1s due to the fact that only a limited number of queries have *been* tested on a single database (we are beginning the implementation phase, following a one year study period).

REFERENCES

<1> Gross. M. Methodes en syntaxe. Paris. Hermann. 1975.

<2> Hoeppner. W.. T. Christaller, H. Marburger, K. Morlk. B. Nebel. M. O'Leary and W. Wahlster. "Beyond Domain-Independence". 1n Proc. of IJCAI-83, Karlsruhe. W Germany, 1983.

<3> Hemphill L. and J. Rhyne, A Model for Knowledge Representation m Natural Language Query Systems, Research Report Laboratory RJ 2304(31046), IBM Research Laboratory, San Jose. September 1978.

<4> Kaplan, S J "Cooperative Responses from a Portable Natural Language Ouery System" Artificial Intelligence, vol 19 (1982) 165-1B7.

<5> Normier. B. SAPHIR - Presentation generale, ERLI. Janvier 1984.

<6> Normier. B.. M-E. Aubert. G Clemencin, B Euzenat . S Lacep, N. Lellouche, A. Ogonowski and J. Vega, Ex tens ion de la gestion du dialogue, Rapport final (Marche DRET 83/234). ERLI. Decembre 1984,

<7> Euzenat. B . , B. Norm ler, A Ogonowsk i and G P . Zarn, Les mecanismes d' inference dans 1 ' interrogation de bases de donnees relat tonnelles en langage nature!, Convention ADI-84-068 rapport intermediate N.1. ERLI. Juin 1984a

<8> Euzenat. B.. B. Normier, A. Ogonowski and G. P. Zarri. Les mecamsmes d'inference dans 1 ' interrogation de bases de donnees relationnel1es en langage naturei. Convention ADI-84-068 rapport final. ERLI. Decembre 1984b

<9> Zarri, G P. "An Outline of the Representation and Use of Temporal Data in the RESEDA System" Information Technology/ Research *and* Development, vol 2 (1983) 89-108

<10> Zarn, G P. "Expert Systems and Information Retrieval: an Experiment in the Domain of Biographical Data Management", in Developments in Expert Systems, Coombs M,J ed. London. Academic Press. 1984a

<11> Zarri. G. P "Constructing and Utilizing a Large Fact Database Using AI Techniques". in Proc of the First International Workshop on Expert Database Systems, Univ of South Carolina. College of Business Administration. Kerschberg L ed . 1984b.

<12> Zarri, G P. "Intelligent Information Retrieval on Interesting Application Area for the New Generation Computer Systems", in Proc of the International Conference on Fifth Generation Computer Systems, Tokyo. Japan. 1984c.