A GENERAL SEMANTIC ANALYSER FOR DATA BASE ACCESS

B. K. Bog & K. Sparck Jones

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

ABSTRACT

The paper discusses the design principles and current status of a natural language front end for access to data bases. This Is based on the use, first, of a semantically-oriented question analyser exploiting general, language-wide semantic categories and patterns, rather than data base-specific ones; and, second, of a data base-oriented translation component for obtaining search specifications from the meaning representations for questions derived by the analyser. This approach is motivated by the desire to reduce the effort of providing data base-specific material for the front end, by the belief that a general analyser is well suited to the "casual" data base user, and by the assumption that the rich semantic apparatus used will be both adequate as a means of analysis and appropriate as a tool for linking the characterisations of input and data language items. The paper describes this approach in more detail, with emphasis on the existing, tested, analyser.

We are trying to build a front end processor for natural-language access to data bases based on the principle that the semantics as well as syntax of the language analyser should be general and not data-specific. The function of the analyser is to build a meaning representation of an input question which is then converted to a formal data base search query, specification by a separate translation component.

The reason for adopting this approach is to reduce the analyser's dependence on the specific properties of the data base, and hence the cost of dealing with new data bases. The properties of the data base necessarily determine the form of search specifications; but the assumption is that the transition from text question to search query is most naturally and effectively achieved by the use of a meaning representation for the input text which is normalised, i.e. reduces input text variation, and which is wholly, or at least primarily, independent of the data base application. Data base-specific information is embodied in translation rules linking the semantic categories and relations of the text meaning representation with those of the data base.

This research is supported by the Science Research Council.

The claim underlying the approach is that the categories and relations of the general meaning representation language are on the one hand powerful enough to capture the sense of the Input question, and on the other can be naturally linked with whatever categories and relations are used to characterise the data, both substantively and administratively. The project differs on the one hand from e.g. PLANES [1] and LADDER [2] in not relying on data base semantics In the analyser, and on the other from e.g. PHLIQA1 [33 in not deferring semantic processing until translation. The analyser is essentially a general semantic processor providing question representations for further, specialised interpretation in the data base context.

The project is based on an existing analyser developing ideas derived from Wilks [4], The analyser exploits a semantic apparatus relying on an extensive system of semantic primitives and types of semantic pattern. Parsing is under the control of an ATN processor, combining the use of conventional syntax for identifying sentence constituents with semantic checks and structure-building actions. The output meaning representation consists of a case-labelled dependency structure with elements defined by semantic primitive formulas.

More specifically, the analyser is designed to Identify noun phrases, clauses, and modifiers (especially prepositional phrases), using an orthodox grammar, currently of moderate extent. These sentence constituents are then interpreted and assembled, and the resulting structure labelled, by satisfying the semantic preferences expressed by the various types of pattern applied. The more important of these patterns are 'templates[1], representing basic propositional message forms; 'pre plates' and 'preps pecs', representing prepositional modifier patterns for free prepositional phrases and post verb constructions respectively; and, most important, 'verb frames', representing the case contexts of verbs. The slots in all of these patterns are defined by the semantic primitive characterisations required of the heads of Individual word formulas, or, in

complex constructions, of the heads of their governing words. The essential idea behind preference semantics is that those matches which satisfy most requirements are selected, even if not all of the specified requirements oan be met.

The approach, following Wilks, is specifically intended to deal with word sense selection, and with semantically-motivated modifier phrase attachment. However the system differs significantly from Wilks* in the use of conventional syntax for constituent identification, in the adoption of the ATM processing mechanism, and in the provision of a more fully structured and explicitly labelled meaning representation. As an illustration, given the question "WHAT IS THE WEIGHT OF THE BLUE PARTS WHICH ARE SUPPLIED BY CURE TO LONDON?", the analyser outputs the representation of Figure 1. In the figure 'be1','supply 1' etc. Identify selected input word senses which are characterised by the Immediately following semantlo primitive formulae, while *##agent' etc. are dependency structure case labels* This representation can be summarily glossed by saying we are querying the ttstate weight(...) of an llagent(many...inanimate things) which (...be) in a flstate oolour(...) and are (...given) by an MagentC ...man) to a ftdestlnatlon(... point).

The system has already demonstrated its capacity to disambiguate lexically and structurally complex sentences (51. However for the purposes of data base access the analyser must be developed: thus mechanisms for treating quantiflcational structures will have to be provided; it is also probable that as the meaning representations derived are not very radically normalised, further processing may be required prior to, rather than during, translation to bring different meaning representations which map onto the same search specification closer together.

The use of general semantic categories and patterns, rather than ones appropriate to particular universes of discourse, rests on what may be called an 'anti-frame' (or 'anti-script') approach to text analysis, or at least on the view that while particular pragmatic knowledge may be relevant to language understanding, it does not play a dominant part in text processing. However rejecting universe of discourse-specific concepts and relationships makes the task of the analyser much more onerous than It is in most data base systems. The use of data base-specific categories and patterns in these systems, and of dictionaries limited to germane senses of lexical items, eliminates a good deal of interpretive effort. In this project lexical items, even if selected for their data base relevance, will not be characterised in terms of their specific data-base readings and contexts.

Unfortunately, this approach leads to acute difficulties with the noun-noun compounds which are

```
Sentence:  WHAT IS THE WEIGHT OF THE BLUE PARTS WHICH ARE SUPPLIED BY CLARK TO LONDON?

(clause (type question) (tns present)
   (v
      (be2
         ((*ent subj) (((own state) obje) be))
         (@@ agent
            ((trace (clause v obj))
               (clause (type relative) (tns present) (aspect (passive))
                  (v
                     (supply1
                        ((*org subj)
                           ((*inan obje)
                              ((*org recipient) (((((@recipient subj) have) cause) goal) give))))
                        (@@ agent (Clark (mal (indiv man))))
                        (@@ obj
                           ((trace (clause v agent))
                              (clause
                                 (v
                                    (be2 ((*ent subj) (((own state) obje) be))
                                       (@@ agent
                                          (part1
                                             ((*inan poss) ((work goal) (subj thing)))
                                             (@@ number many)))
                                       (@@ state colour
                                          (val
                                             (blue1
                                                ((*inan poss)
                                                   (((man subj) (see sense)) (obje kind)))) ))) ))) )
                        (@@ destination (London (this (where point)))) ))) ))
         (@@ state weight (val (query (dummy)))) )))
```

FIGURE 1

especially rampant in data base worlds, for example ♦•apartment block building regulations" or "wing flap elevator control parts". Since semantic interpretation and description is interleaved with syntactic processing in the operation of the analyser, if specialised pragmatic knowledge is required to handle such constructions, this would seem to imply that the attempt to build an analyser using only general, language-wide, knowledge is doomed to failure. However we make the distinction between basing the analyser primarily and explicitly on data-base specific semantics, and basing it on general semantics with provision for the invocation of supporting particular knowledge. The project is essentially an experiment designed to show whether this distinction can be maintained. Thus the intention is that the relevant specific knowledge, in the form in which it is required for the translation component, should also be available for use, in a controlled and sufficiently universal fashion, by the analyser.

For initial tests a relational data base will be used, with search specifications in relational calculus or algebra form. It is not clear whether it will be possible or appropriate to treat the translation component just as a set of production rules mapping meaning representation structures into search specifications, in the style of LUNAR, or, more elaborately, as a set of procedures exploiting an enriched data base characterisation in e.g. the form of a semantic network. In either case, however, the assumption is that It will be possible to link the language of the input questions with that of the data base by a common characterisation of concepts and relationships in terms of the semantic primitives so far used in dealing with text inputs.

For example, if we hypothesise a data base dealing, in familiar style, with suppliers and parts, we will have formulas for words occurring in questions like "PART" or "SUPPLIER", and meaning representations for phrases, like "BLUE PARTS" (cf Figure 1). The data base words used in the data description, i.e. the data base entity, attribute and relation type names, will be similarly characterised, and this common method of meaning characterisation should in principle permit the translation component to connect the elements of the input question and the search query. Of course it cannot be assumed that input expressions corresponding to data base values can be characterised in advance: this is conspicuously a problem with proper names but can occur with other words. The oonnection here between input and data base has to be via common components of input word representations, which should serve to identify appropriate search areas in the data base. Thus the representation for "BLUE" would overlap with that for COLOUR via such a common sub-formula as ((man subjMsee sense)). The semantic primitives of the question meaning representation, and especially the case labels, should similarly promote the translation of the question structure into an appropriate search query form. Thus, to take a very simple example,

66agent(...thing) (...be) f#state(...kind)

will map into a data language expression of the form ENTITY HAVE ATTRIBUTE. For more complex questions the dependency structure of the meaning representation should be easier to manipulate to derive the search query structure than the input text surface syntax. Work recently began on developing a translator along these lines.

REFERENCES

[1] Waltz, D. L. "An English Language Question Answering System for a Large Relational Data Base." Communications of the ACM 21 (1978) 526-539.

[2] Hendrix, D. G. et al. "Developing a Natural Language Interface to Complex Data." ACM Transactions on Database Systems 3 (1978) 105-147.

[3] Bronnenberg, W. J. et al. "The Question Answering System PHLIQA1." In Natural Language Answering Systems (Ed. Bole)♦ London: Mac roil lan, 1979.

[4] Wilks, Y. "A Preferential, Pattern-seeking Semantics for Natural Language Inference." Artificial Intelligence 6 (1975) 53-7**.

[5] Boguraev, B. K. "Automatic Resolution of Linguistic Ambiguities", TR-11, Computer Laboratory, University of Cambridge, Cambridge, England, 1979.