COLING 82, J. Horecký (ed.)
North-Holland Publishing Company
© Academia, 1982

NATURAL LANGUAGE INTERFACES USING LIMITED SEMANTIC INFORMATION

Ralph Grishman, Lynette Hirschman*, and Carol Friedman

New York University
New York, NY

In order to analyze their input properly, natural language interfaces require access to domain-specific semantic information. However, design considerations for practical systems -- in particular, the desire to construct interfaces which are readily portable to new domains -- require us to limit and segregate this domain-specific information. We consider here the possibility of limiting ourselves to a characterization of the _structure of information_ in a domain. This structure is captured in a _domain information schema_, which specifies the semantic classes of the domain, the words and phrases which belong to these classes, and the predicate-argument relationships among members of these classes which are meaningful in the domain. We describe how this schema is used by the various stages of two large natural language processing systems.

The necessity of incorporating domain-specific semantic information into natural language processing systems is now generally recognized. The task we face as computational linguists lies in selecting this information, organizing it, and integrating it into a natural language processing system.

In principle, no limit can be placed on the semantic knowledge needed for natural language analysis -- given essentially any fact, one can devise a natural language input which requires knowledge of that fact for its correct interpretation. For the construction of operational systems, however, there are practical limitations on our ability to collect and organize the domain-specific knowledge for any substantial domain. Rather than ignore such limitations, we should use them as a motivation for identifying manageable components of this domain-specific knowledge. Such considerations are especially important if we are aiming to construct _portable_ systems -- systems which can be readily moved from one domain to another.

What properties should such a component have? It should

* be _effective_ in providing the information needed to guide the analysis of the input text;

* have a _simple structure_, to facilitate both the collection of the information and its use in the language analysis procedures;

* have a _discovery procedure_ -- a systematic way of collecting

---

* Present affiliation: Research and Development Activity, Federal and Special Systems Group, Burroughs Corp., Paoli, PA.

this information for a new domain.

We suggest that a characterization of the structure of information in a domain is such a semantic component. We call this component a domain information schema (DIS). A DIS specifies a set of semantic classes, the words and phrases which belong to these classes, and the predicate-argument relationships among members of these classes which are meaningful in this domain. Some features of these relationships, such as functional dependencies between semantic classes, are also noted.

This is not a novel assemblage of information. The DIS is perhaps most similar to data base schemata which also seek to separate a description of the structure of information in a domain from the specific facts about a domain. In frame-based systems, this information is essentially captured by the top-level frames, although the delineation here between structural description and specific facts is not as precise. Semantic grammars embed much of the information of the DIS, although there it is mixed with general linguistic knowledge. Certain parsers (e.g., the RUS parser [1]) also make use of aspects of information stored in a separate semantic component. Thus information similar to a DIS has been used, at least implicitly, by other natural language systems; however, little research has been explicitly concerned with the task of choosing a subset of the domain-specific information and evaluating it using criteria such as those mentioned above. We therefore decided to address this question with respect to the DIS in our recent research.

To this end, we have recently modified portions of two large natural language systems so that all domain-specific knowledge is isolated in a DIS. One of these is a system for the information formatting of natural language medical reports; the other, a "question-answering" system for data base retrieval using natural language. We shall report here on how information from the DIS is used in the various stages of analysis.*

THE SYSTEMS

The information formatting system [2] is designed to accept natural language text in some scientific or technical domain and map the text into a domain-specific structure (an information format) which is suitable for subsequent retrieval operations. In essence, the format is a set of tables in which each category of domain information (for example, for hospital reports: laboratory tests, laboratory findings, diagnoses, treatments, etc.) is assigned a separate column. This formatting procedure has been successfully applied to radiology reports and to hospital discharge summaries. The question-answering system [3] accepts natural language queries regarding the data in the text and retrieves the requested information from the formatted data base.

Both systems use the Linguistic String Parser and grammar [4] to obtain a parse and transformational decomposition of the input sentence. The grammar is an augmented context-free grammar written in Restriction Language [5]. In the formatting procedure, the

* We have concurrently been investigating discovery procedures for DIS's; some of our early work in this area was reported in [6].

decomposition tree is mapped into the information format; the format then goes through a normalization component which fills in implicit information and a component to analyze the time structure of the narrative. For question answering, the decomposition tree is mapped into an extended predicate calculus formula; this is followed by anaphora resolution and translation of the formula into a data base retrieval request.

SELECTION

The domain information schema is most directly reflected in the syntax of the language, forming a sublanguage as described by Harris [7]. The semantic classes and relationships, as defined by the DIS, are used to formulate sublanguage selectional constraints. These constraints rule out incorrect syntactic analyses, many of which are caused by structural ambiguity due to adjunct placement and conjunction, and by lexical ambiguity due to homographs.

The selection mechanism is list driven to provide for portability from one sublanguage to another. These lists specify for each basic linguistic relation, such as SUBJECT-VERB-OBJECT or HOST-ADJECTIVE, the patterns of word classes which are permissible in the sublanguage. Each basic lingustic relation has many surface realizations for which selection must be checked. The SUBJECT-VERB-OBJECT relation, for instance, may appear in declaratives and questions, in main and relative clauses, in active and passive voice, in perfect and progressive forms, etc. This task is greatly simplified, however, by the linguistic routines of the Restriction Language [4,5], which locate the elements of the parse tree bearing the underlying SUBJECT-VERB, VERB-OBJECT, and HOST-ADJUNCT relations.

An example of how the DIS eliminates incorrect parses in the medical sublanguage can be seen in the sentence from a medical text

Brother 18 also has heart disease, on cardiac meds.

which has two analyses: one where "on cardiac meds" is an adjunct of "heart disease" and the other where it is an adjunct of "brother". There is a HOST-ADJUNCT pattern for the classes FAMILY-MEMBER ON MEDICATION but not for DIAGNOSIS ON MEDICATION; thus only the second analysis has a pattern matching one in the DIS.

Matching the patterns is only one function of the selection procedure. When a match is successful, those classes which match the pattern are recorded as "selected attributes" so that they may be referenced at a further point in processing. Once a pattern is established, the "selected attribute" classes are preferred to the original ones. Additional selectional constraints will refer to the "selected attributes" of a word if it exists. How this procedure aides in the disambiguation of homographs can be shown using the homograph "discharge". "Discharge" can be a medical administrative action (MED-VERB) as in "discharge from hospital" or a SIGN-SYMPTOM word as in "discharge from wound". The phrase "discharge from hospital" will be successfully matched by the pattern MED-VERB FROM INSTITUTION; there is, in contrast, no pattern SIGN-SYMPTOM FROM INSTITUTION. Thus in this phrase "discharge" is assigned a "selected attribute" MED-VERB and the

SIGN-SYMPTOM class of "discharge" will be ignored. This will be particularly helpful in the information formatting stage, since the mapping into the format is based primarily on a word's selected sublanguage class.

The selectional constraints are complicated by the fact that the class of a noun phrase is sometimes determined by the entire phrase and not by the head noun alone. In some cases the class of the phrase is the class of one of its constituents. For example, "stiff neck" has the same class as "stiff", which is a SIGN-SYMPTOM class. In other cases words from two classes combine to form a phrase with a different class. In the medical domain, "temperature of 103" is of the FINDING class because "temperature" is in the BODY-FUNCTION class and "103" is a quantifier. This computation of a phrasal attribute is called the "computed attribute" construction. This attribute plays an important role in eliminating incorrect parses which arise with coordinate conjunction. Noun phrase conjunction is restricted to phrases which are of the same or closely related classes. In "Patient had stiff neck and fever" there are two readings. The reading in which "stiff" is the left adjunct of both "neck" and "fever" is eliminated because "neck" and "fever" have different subclasses: "fever" is a SIGN-SYMPTOM word whereas "neck" is a BODY-PART word. However the phrase "stiff neck" has a SIGN-SYMPTOM "computed attribute" and is in the same class as "fever"; therefore we do get the analysis where "fever" is conjoined to "stiff neck". A more detailed description of constraints on noun phrase conjunction is described by Hirschman [8].

FORMATTING

The format itself can be viewed as a derivative of the DIS, obtained by merging several predicate-argument relations into a single larger relation. Because the formats, like the predicate-argument relations, are based on the semantic classes of the DIS, the mapping from decomposition trees into formats can be driven by a table of the correspondences between semantic classes and format columns.

QUESTION-ANSWERING

The predicate names used in the predicate calculus representation within the question-answering system correspond to the predicate-argument patterns of semantic classes in the DIS, so the mapping from decomposition trees to predicate calculus expressions is also DIS-driven. In addition, this mapping uses the information on functional dependencies recorded in the DIS: quantifier scoping is determined primarily by surface word order and syntactic structure, but functional dependencies may take precedence. For example, in the medical domain, because there is a functional relation from "X-rays" to "patients" (each X-ray is of one and only one patient), the phrase "the X-rays of the patients" is correctly analyzed with the quantifier over "patients" having wider scope than the quantifier over "X-rays".

The anaphora resolution component relies on the selection mechanism described earlier (and hence on the DIS) to determine from context the possible semantic classes for the referent of an anaphoric phrase; the antecedent search is then restricted to members of these classes. In addition, the word classes are used in distinguishing between definite and "one" anaphora (as defined by Webber [9]), and resolving "one" anaphora correctly [10].

CONCLUSION

In summary, the DIS has proven in these systems to be an effective source of domain-specific information. System portability has been enhanced by using information of simple structure which can be isolated from the linguistic processing mechanisms. At the same time, the simplicity of structure has facilitated the integration of this information into many stages of the analysis procedure.

ACKNOWLEDGEMENTS

This research was supported in part by National Science Foundation grants MCS 80-02453 from the Division of Mathematical and Computer Sciences and IST 81-15669 from the Division of Information Science and Technology; in part by National Library of Medicine grant 1-R01-LM03933, awarded by the National Institutes of Health, Department of Health and Human Services; and in part by Office of Naval Research contract N00014-75-C-0571.

REFERENCES

[1] Bobrow, R.J. and Webber, B.L. Knowledge Representation for Syntactic/Semantic Processing, First Annual Nat'l Conf. on Artificial Intelligence, 316-323, AAAI, Stanford, 1980.

[2] Sager, N. Natural Language Information Formatting: The Automatic Conversion of Texts to a Structured Data Base. In Advances in Computers 17 (M.C. Yovits, ed.), 89-162 (Academic Press, NY, 1978).

[3] Grishman, R., and Hirschman, L. Question Answering from Natural Language Medical Data Bases. Artificial Intelligence 11 (1978), 25-43.

[4] Sager, N. Natural Language Information Processing (Addison-Wesley, 1981).

[5] Sager, N., and Grishman, R. The restriction language for computer grammars of natural language. Comm. ACM 18, 7 (July 1975), 390-400.

[6] Hirschman, L., Grishman, R., and Sager, N. Grammatically-Based Automatic Word Class Formation. Information Processing and Management 11 (1975), 39-57.

[7] Harris, Z. Mathematical Structures of Language (Interscience, New York, 1968).

[8] Hirschman, L. Constraints on Noun Phrase Conjunction: a Domain-Independent Mechanism. Proc. COLING 82 (this volume).

[9] Webber, B. A Formal Approach to Discourse Anaphora (Garland, New York, 1979).

[10] Grishman, R. Resolving Noun Phrase Anaphora. Paper presented at the Assn. for Computational Linguistics meeting on "Computer Modeling of Linguistic Theory," New York, Dec. 28, 1981.