

A TWO LEVEL MODULAR SYSTEM FOR NATURAL LANGUAGE UNDERSTANDING

Giovanni Guida and Marco Somalvico
Milan Polytechnic Artificial Intelligence Project
Istituto di Elettrotecnica ed Elettronica
Politecnico di Milano
Piazza Leonardo da Vinci n. 32
1-20133 Milano, Italy

This paper is intended to propose a new methodological approach to the conception of natural language understanding systems. This contribution is supported by the design and implementation of DONAU. The system is based on a two level modular architecture. The horizontal level allows an independent and parallel development of the single segments of the system (syntactic analyser, semantic analyser, information extractor, legality controller). The vertical level ensures the possibility of changing the definition of the semantic domain on which particular versions of the system can be oriented and specialized in a simple, incremental, and user-oriented way. DONAU is written in LISP and is available from the authors.

1. INTRODUCTION

The conception of modular and flexible architectures has been proposed recently in the development of natural language understanding systems [2]. Basic exigencies of unification, flexibility, and intercommunication have motivated the development of the DONAU (Domain Oriented Natural language Understanding) system, which has been successfully experimented on an UNIVAC 1110 at the Milan Polytechnic Artificial Intelligence Project [3].

The original contribution of DONAU consists in the adoption of a two level modular architecture for a system able of understanding natural language.

In the development of the DONAU two level architecture, the morphologic and syntactic analysis has been designed and implemented on the base of the PIAF system which has been conceived, developed, and successfully tested in the past at the IMAG Laboratory of the University of Grenoble (France) [4].

Two versions of DONAU have been already completely implemented and experimented: a first one, oriented on robotics, for the programming and control of the SUPERSIGMA robot developed and operating at the Milan Polytechnic Artificial Intelligence Project; a second one, oriented on data bases, for the inquiry of a relational data base about the oil resources in the world proposed by the International Institute for Applied System Analysis, Laxenburg, Austria. A third version, centered on program synthesis, is presently being designed [1].

2. THE TWO LEVEL MODULAR STRUCTURE

This section illustrates the modular structure of DONAU and the adopted design criteria. The main criterion which has guided the conception of DONAU has been that one of two level modularity.

The first level, called horizontal level, reflects the distinction between the different conceptual activities concerning the natural language understanding process. Eachone of these activities (i.e., morphologic and syntactic analysis, semantic analysis, information extraction, and legality control) is associated with a bimodular segment which is precisely defined by its input-output characteristics. The second level of the modular architecture, called vertical level, is related with the distinction of the different nature of the two modules which compose the bimodular segment. The first module, called top module, embeds the domain-independent, general-purpose algorithm associated with the conceptual activity proper of the segment. The second module, called bottom module, consists of a framework containing the domain-dependent, special-purpose information which directly relates with the particular semantic domain in which a version of the DONAU system is conceived to be operating.

An advantage of this two level architecture lies in an efficient balance between the characteristic of generality, required by a natural language understanding system, and the exigency of efficiency, implied by an actually useful system.

Another feature is the simplicity in designing different versions of the same system, each one devoted to a particular instance of a semantic domain. Each new version can be obtained, in fact, by simply changing, in each segment, only the bottom module, i.e., the domain oriented information. A further quality is the possibility of having incrementality in the definition and set up of the system, i.e., the capability of modifying and increasing the content of the specification information on the base of the requirements and suggestions which the experimentation of the system might propose.

The fundamental assumption on which the DONAU project is based is that one of artificial understanding of natural language. We claim that the understanding of natural language, having as ultimate goal the communication between the man and an artificial system, does not require the capability of following all the peculiarities and the nuances of the possibly different ways of expressing an operative information in natural language. The goal of the understanding process is to obtain just that kernel of information which is necessary to activate in the desired and correct way the artificial system which is the target of the man-machine interaction and which can perform only a bounded, usually quite small, number of activities. This approach to natural language understanding can be useful, in our opinion, for avoiding some of the linguistic difficulties encountered in early research works. Moreover, this standpoint allows to largely extend the capabilities of the system in accepting a great variety of free natural language sentences, still avoiding the explosion of the complexity of the parsing algorithm consequent to an increased complexity of the input sentences which might actually be encountered in real applications.

3. UNDERSTANDING INPUT SENTENCES

The first step of the parsing consists of the morphological and syntactic analysis. The classical methods of finite state recognizers and dependence grammars [4] are adopted for this activity. The result of the syntactic analysis, constituted by a (small) set of dependence trees [3], is then supplied as input of the semantic analyser.

The role of the semantic analysis is to discriminate, among the different dependence trees, that one which is consistent with the semantic description of the natural language supplied to the system. The specification information of this segment is constituted by a complex framework called model-list network. This network is composed of an hierarchical organization of model-

lists. The model-lists are lists of symbols representing morphological types and semantic types, which denote the categories of words and elementary constructs considered as relevant ones for defining the semantics of the natural language in a particular application domain. Each model-list is associated to a semantic type and constitute therefore a transformation rule. A dependence tree is discriminated by the semantic analyser if to its nonterminal nodes model-lists can be associated in such a way to satisfy an appropriate set of matching conditions [3]. The output of the semantic analyser is represented by a semantic tree embedding all the relevant information so far obtained in the parsing.

The third segment of DONAU is the information extractor. Its role is to individuate within the semantic tree the operative significant content which is sufficient and not redundant for expressing the meaning of the original natural language sentence. The elimination of the superfluous information is performed by a covering algorithm on the base of the content diagram [3] defined on the lexicon. The result of the whole parsing activity is an executable formal statement which can directly be accepted and executed by the interacting artificial system.

Finally, the legality control consists in the evaluation of the proposed operative content within a legality base which represents the non-linguistic context of the artificial system connected with DONAU.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

Two versions of DONAU have been developed.

RB-DONAU, centered on the semantic domain of robotics, is running, at present time, on an UNIVAC 1110 computer and is programmed in 1108-LISP. The software architecture of the system reflects quite closely the conceptual schema described in section 2. The program occupies about 25 Kwords memory and the specification information, relative to a lexicon of about 200 words, 5 kwords. The comprehension of simple sentences of 5 to 10 words requires about 1 second (with compiled LISP programs).

DB-DONAU, centered on the domain of data base inquiry [1], has been developed with the precise aim of verifying the actual possibility of changing the semantic domain along the lines suggested by the vertical modularity of the system. The experimental activity done has shown a good flexibility of the model proposed for achieving the modularity and incrementality goals initially stated for the project.

Several projects are presently developing around the DONAU system. In particular, the following goals are considered as relevant ones :

- to improve the diagnostic capabilities of DONAU and to develop an high-level testing and tracing subsystem;
- to define and implement an experience subsystem, which can guide the user's tuning and extension activities on the base of the past use of the system;
- to integrate the two steps of syntactic and semantic analysis as parallel coroutines for increasing the efficiency of the parser.

APPENDIX

In this appendix we present an example of natural language understanding on a simple sentence :

ELENCA I DEPOSITI DEL KUWAIT AVENTI RISERVA DI PETROLIO MAGGIORE DI 1500 TON AL 1.9.1978 (list the deposits of Kuwait having an oil reserve greater than 1500 ton on 1.9.1978).

Two different dependence trees are obtained:

```
((ELENCA verb)*((DEPOSITI sub) (I art)*
((KUWAIT sub) (del prep)*((AVENTI verb)*
((RISERVA sub)*((PETROLIO sub) (DI prep)*
(MAGGIORE adj) ((1500 num) (DI prep)*
(TON sub)) ((AL prep)* (1.9.1978 date))))))
```

```
((ELENCA verb)*((DEPOSITI sub) (I art)*
((KUWAIT sub) (DEL prep)*
((RISERVA sub) (AVENTI verb)*((PETROLIO sub)
(DI prep)* (MAGGIORE adj) ((1500 num) (DI prep)*
(TON sub)) ((AL prep)* (1.9.1978 date))))
```

The notation (x y * z) is used, where x is the root of the tree, y are the left subtrees and z are the right subtrees.

Only the first tree is discriminated by the semantic analyzer; the corresponding semantic tree is :

```
((sentence: (ELENCA verb))*((object: (DEPOSITI
sub)) (I art)*((spec: (KUWAIT sub)) (DEL prep)*
((compl: (AVENTI verb))*((compl-spec: (RISERVA
sub))*((spec: (PETROLIO sub)) (DI prep)*
(MAGGIORE adj) ((spec: (1500 num)) (DI prep)*
(TON sub)) ((date: (AL prep)* (1.9.1978 date))))))
```

The resulting information tree is then:

```
((sentence: (ELENCA verb))*((object: (DEPOSITI
sub))*((spec: (KUWAIT sub)) ((compl:)*
((compl-spec: (RISERVA sub))*((MAGGIORE adj)
((spec: (1500 num))* (TON sub)) ((date:)*
(1.9.1978 date))))))
```

ACKNOWLEDGMENTS

We are grateful to the researchers of the IMAG

Laboratory (University of Grenoble, France) which have cooperated to this project and to the students of the Politecnico di Milano for their contributions in the experimental work.

REFERENCES

- [1] Bernorio, M., Bertoni, M., Dabbene, A., Somalvico, M. "A Domain Oriented Natural Language Understanding System for Man-machine Interaction with Dynamic DATA Bases". In Proc. IIASA Workshop Natural Languages for Interaction with Data Bases, Laxenburg, Austria, October, 1978, pp.239-264.
- [2] Hendrix, G.G. "A Natural Language Interface Facility and its Application to a Data Base" In Proc. IIASA Workshop Natural Language for Interaction with Data Bases, Laxenburg, Austria, October, 1978, pp.87-94.
- [3] Guida, G., Somalvico, M. "DONAU: A General Purpose Domain Oriented Natural Language Understanding System". Memo MP-AIM-99, Politecnico di Milano, Milan, Italy, October, 1978.
- [4] Veillon, G. "Modèles et algorithmes pour la Traduction Automatique". Thèse, Université Scientifique et Médicale de Grenoble, Grenoble, France, 1970.
- [5] Woods, W.A. "Transition Network Grammars for Natural Language Analysis", Comm. ACM 13:10 (1970) 591-606.