ONGOING RESEARCH IN COMPUTATIONAL LINGUISTICS AT THE UNIVERSITY OF ANTWERP (UIA)

Luc Steels
UIA, Dept. Ger.Fil
Universiteitsplein,1
B-2610 Wilrijk-Antwerpen

There exist two basic types of research in computational linguistics. The first type is closely linked to statistical research on language. The computer is here used to cope with the problem of investigating a large corpus in a reasonable time span. The other type is of a rather different sort. Here the computer is used as an instrument to simulate the extremely complex process of natural language understanding and producing. Recently several remarkable results have been obtained in this area. We mention the creation of question/answering systems, such as Woods'(1972) system that could answer questions stated in natural language about the NASA data base on moon rocks, or of other man-machine interaction systems, such as Winograd's famous program where a computerprogram performs physieal actions of manipulating blocks after being told in natural language what to do(Winograd(1972)). Other problems that are approached are machine translation, speech understanding, speech synthesis, etc;. We call the second type of research in computational linguistics natural language processing.

The research currently going on at the UIA is exclusively directed to natural language processing. This is partly due to the type of machinery that we have available at the moment, partly because we feel that this is the more exciting part of the discipline. In this paper, some indications are given about what areas of natural language processing we are investigating and what partial results could already be obtained. For a general introduction into the problems and methods of natural language processing research, we refer to Steels (1976a).

First we want to clear up a common misunderstanding about what is exactly involved in natural language processing. (Too)many linguists, even today, think that the construction of computational systems for linguistic tasks such as language analysis and synthesis is a basically technical problem. This is wrong. Of course the construction of such systems does require knowledge about computer programming and related aspects, but the main problem in the development of an explicit, exact and empirically sound theory of natural language. And if such a theory would have been

created by linguistics proper, the activity of the computational linguist would be reduced to translating the formalism of the linguistic theory into that of computer programs. Unfortunately (or should we say fortunately) this is not the case. The linguistic theories existing today are far from adequate as regards the description of linguistic systems. As a consequence research in natural language processing is as much (or even more) concentrating on the development of new linguistic theories, than on the actual design and implementation of applications. In general, it is felt that we are not yet far enough to construct systems for use in the real world anyway.

Now we start discussing very shortly the kind of research being undertaken at the moment in our computational linguistics laboratory.

1. Software

The first sort of problems we are dealing with are of a mere technical nature. We are developing a library of functions and subroutines (written in the programming language FORTRAN IV) which is of relevance for linguistics. This includes string manipulation, list processing, internal tree representation, the plotting of trees, routines for processing semantic networks, etc;. The library is constantly growing to meet the needs of the researchers. Aspects of it are documented in Steels (1976b).

2. Linguistic systems

From January 1976 until June 1976, a natural language producing system was designed and implemented with the following properties:

(i) The system takes as input a language free semantic representation and produces a natural language sentence.

(ii) Any language can be dealt with because the language specific information was consulted as data external to the program itself. (Experiments were carried out for Dutch, English, French and German).

(iii) Theoretically, the production system is based on the interaction of several knowledge sources: word order, categorial patterns, syntactic features concord, semantic features (selection restrictions matches), case frames, in particular surface case signals, morphological processes.

Documentation: A first report on the overall system was presented at the computational linguistics conference in Ottawa (July 1976). Other reports about the formal and mathematical basis of the approach and other aspects were published in the literature.

Evaluation: The experiment was considered a success, although for real world application a second experimental version and maybe even a third one should be designed and implemented.

From June 1976 until December 1976 a first experimental natural language analyzer was designed and implemented. This system takes a natural language sentence as input and returns the semantic structure for it as output. Again any language can in principle be dealt with because the language specific information (i.e. the grammar) is consulted as data external to the system itself. The consulted information contains knowledge about word order, categorial patters, syntactic features concord, semantic features (selection restrictions matches), case frames tests, in particular surface case signals matches, a.o.. An effort has been made to let the linguistic information be the same both in the parsing and production system.

Documentation: A video-film has been prepared about the analysis process. It was first presented at a colloquium on computational linguistics (UIA, October 1976). Several reports are forthcoming about the principles of the parser.

Evaluation: Although several problems remain (e.g. the processing of coordination), the parser is doing very well. But again for use in real world applications, other experimental versions must be developed. There is also work being undertaken by members of the laboratory on semantic interpreter systems to be coupled to the parser.

It is important to note that the systems discussed here are in no way reconstructions of systems used in other computational linguistics laboratories. Instead, fundamental research has been undertaken to develop new types of language processing systems and new linguistic theories. This includes both formal investigations of the theories being used and empirical language studies.

Future work: Besides further work on the existing systems and especially research in computational semantics for natural languages, several other systems are under development at the moment. The most important ones are (i) a question/answering system for family relations, with as query language a Dutch-like language (ii) a system embedded in

the parser simulating 'literary' aspects of language understanding such as metaphor, plot perceptions, a.o.. and (iii) a system for the automatic translation of notes on catalog cards in libraries. We discuss in some more detail the goals of the latter system.

On each catalog card in a library occurs one (optional) field called the annotation field which is used to express information about the document not expressable in the author-,title-,editor-,etc;-fields. Such an annotation field has two important characteristics. (1) It is stated in the local language of the users of the library, (2) a great amount of freedom is involved from the part of the catalogue writer.
When catalogue cards are exchanged among libraries of different language communities, and they are, there is the problem (due to (1)) that the annotation field must be translated into the local language. However due to (2) this translation process is essentially nontrivial. That means algorithms which do not take the processing of natural language serious will necessarily fail. This fact turns out to be a serious problem on our way to a full mechanization of catalogue cards exchange.
To solve the problem we are in the process of constructing an automatic translation system that takes a note in whatever language and translates the note in the language of the user. Such a system is expected to become operational in June 1977.

In this paper we presented briefly what kind of investigations we are undertaking in the area of natural language processing at the moment. The most important work is on typically linguistic systems for parsing and producing natural language and for semantic processing. More details about all aspects can be obtained from the author.

References

Steels,L. (1976a) Introduction to natural language processing. In: Steels, L. (ed.) Advances in natural language processing. Preprints of a workshop, UIA, October 1976.

Steels, L. (1976b) The FORLI.OLB package for list processing in FORTRAN IV. A user's manual. Antwerp papers in linguistics, nr. 9.

Winograd, T. (1972) Understanding natural language. University of Edinburgh Press. Edinburgh.

Woods,W.,et.al. (1972) The Lunar Sciences Natural Language Information System. Final Report. BBN report N. 2378.