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

TULIPS-2 - NATURAL LANGUAGE LEARNING SYSTEM

Michael G. Malkovsky
Computational mathematics and Cybernetics Faculty
Moscow State University
Moscow
U.S.S.R.

The learning of a natural language is considered to be an important aspect of man-machine communication in human language. The methods of the Russian language knowledge representation and acquisition implemented in the experimental understanding system TULIPS-2 are described. These methods provides for understanding utterances that contain words and structures unknown to the system wherther they are grammatical or erroneous items, or the user's speech peculiarities.

In recent years the problem of man-machine communication by means of natural language (NL) is becoming a practical one. And the designers of "large" applied systems have to overcome new difficulties in order to make such a communication a reality and to enable the user to interact with the computer without any special training and much effort, in a way which is convinient to him, but not to the computer.

We think, that a so-called "restricted NL" is a mere fiction of mind. This term has been invented to denote a language used in a certain problem-domain and based on a NL with great restrictions on its structure. In this case it would be more natural to use a special formal language, which can be quickly learned by the user and efficiently implemented. On the other hand to learn the lexicon and the grammar of the restricted NL and above all to follow these restrictions during a continuous dialogue with the system is rather difficult for a human being.

If NL is really needed, the user should be offered the conditions of communication similar (from the information processing point of view) to those in everyday discourse. Firstly, the restrictions, if any, should be minimized and naturally determined by the problem-domain and by the nature of tasks. Secondly, the "listener" of the user should be able to understand the user correctly in a certain situation even if the utterance is potentially ambigious, incomplete, deviant or if it contains words and syntactic structures unknown to the system whether they are grammatical or erroneous. We contend that it is necessary to consider the deviation from the language norms and other speech peculiarities of the user. Thirdly, it should be possible for the system or for the user to suspend their conversation in order to ask the partner a question or give him some advice. However, the specifying dialogue should not occur very often and "on trifles". Finally, the system - like its human partner - should be able not only to act in an unknown situation but to acquire more knowledge, i.e. to learn the language and the user's speech peculiarities.

The experimental system TULIPS (Malkovsky (1975)) and its new version TULIPS-2 (Malkovsky and Volkova (1981)) both were designed in consideration of the above-mentioned demands.

The AI system TULIPS-2 implemented in PLANNER for the BESM-6 computer is intended for further experiments in the field of the computer understanding of NL and for practical use. The system can help the user to form the conditions of a problem. In this case the user gives the system the unformalized description of the problem situation, whereas the system helps to specify this description and to find an adequate formal representation. Such a flexible dialogue using vague terms and loose concepts can be conviniently performed just in a NL (Russian - for TULIPS-2). Moreover the TULIPS-2 system can work in problem-domains with various structures and degrees of formalization. That is another argument for the use of NL.

A user's interaction with the system (via a terminal) is composed of several seances. At the begining of each seance the user have to identify himself and to indicate the problem-domain. This information guides the "tuning" of the system for the seance, i.e. fetching the relevant data from the external memory. This helps to reduce data used in conversation. On the other hand the tuning process introduces the user's speech peculiarities and specific NL items of the problem-domain. During the analysis of utterances these peculiarities and items are looked through before all the other data (lexical, syntactic, and semantic).

Besides, there are the following methods of data representation and handling in the system: special tags define the measure of preferability of relevant data items and procedures and influence the order of their choise during analysis; the lexical items and the grammar rules contain the references to procedures that can be invoked when an item or rule is being handled; NL meta-level items describe the means and range of the Russian language rules alternation by the system; NL knowledge of the system includes both basic knowledge of the Russian language and "open" set of Russian grammar rules, Russian lexical items etc., that can be widened in a seance by the user or by the system itself ("self-taeching").

It should be noted that the basic knowledge is formed and input into the system by its authors or by its operators beforehand. Thus in a seance the system starts to learn NL, to acquire user's speech peculiarities, new terms and abbreviations having much knowledge of NL which make it possible for the system to act in unknown situations by itself. However, change of basic knowledge can be done only with user's permission.

The methods of representation and handling of NL knowledge are important to the system's analyzer which provides for the input message understanding from the context of the conversation. Syntactic, semantic, and pragmatic predictions are widely used on different levels of analysis. The predictions generated from context make it possible to attribute the expected (predicted) characteristics to unknown units, while the references to procedural elements provide for a flexible control, i.e. the possibility of passing on to a more informative (where predictions are more definite) level of analysis.

If necessary the analyzer appeales to the meta-level knowledge - invokes procedures which handle unknown units (words or phrases). These procedures classify such a unit (erroneous form of a known unit or an unknown correct unit) and prepare the information of a unit or an error for storing. The stored information is available both in this seance and in the subsiquent ones.

Sometimes a deviant form can be passed on to further higher levels of analysis, as e.g. the module of spelling correction does. This module processes errors typical for the user working at the terminal (the missing, duplication, permutation of letters or an incorrect shift). However, usually as the result of learning (self-teaching or teaching by user) new items are formed and the old items are changed. The following item types are formed and changed: NL words and phrases descriptions - lexical items and grammar rules, NL meta-level items, control structures - tags and procedures (e.g. special patterns for frequent and typical phrases).

The methods of learning on morphological and lexical levels of Russian have been used in the TULIPS-2 system since 1980. The basic knowledge for these levels includes: a complete description of Russian inflexion, a description of some rules of Russian word-formation and of different typical mistakes made by users, a vocabulary of about 1000 stems, and vocabularies of affixes.

REFERENCES

1 Malkovsky, M.G., TULIPS - Teachable, Understanding Natural Language Problem-Solver, in Proc. of the 4th IJCAI (Tbilisi, 1975).

2 Malkovsky, M.G. and Volkova, I.A., TULIPS-2 Analyzer. Morphological level, Vestnik Moskovskogo Universiteta, Series XV, N 1 (1981) 70-76.