TULIPS - TEACHABLE, UNDERSTANDING NATURAL LANGUAGE INPUT PROBLEM-SOLVER

Michael G. Malkovsky

Computational Mathematics and Cybernetics Department
Moscow State University
Moscow 117234, USSR

Abstract

This paper describes the principal features of TULIPS program designed as a widely-oriented AI system that accepts natural language (NL) input. NL processor of TULIPS is guided by information represented in models of current "external world" domain and of current user and utilizes deductive and inductive mechanisms. These features allow the program: to discover the most relevant interpretation of an input, to extract the descriptions of user's goals from NL utterances, and to generate the so-called T-problems (T for TULIPS). Having solved them TULIPS not only answers user's request but also learns new knowledge on its environment and improves its own mechanisms.

Introduction

It need hardly be said about attractiveness of NL communications with a program. TULIPS's user must not learn special interactive languages and can pass to a computer not a formali»ed description of the algorithm but a description of the problem itself (an input for problem-solving block of the program)• It must be especially emphasised that NL description of a problem can be of the preliminary informal kind. For TULIPS is implemented as a dialog-system, user is able to specify and supplement this description in case of need.

Thus TULIPS cardinally differs from "traditional" computer program. It is not a detailed formal description of an algorithm for the current user's problem, but a generator of such algorithms. It receives as input not a command to activate known to user Btrict segment of algorithm, but only a description of user's problem (i.e. user sais what he wants without any specification how it Bhould be done). Just this property of TULIPS - its ability automatically solve the problem, being told the wording of it - allows to consider the program as AI system that automatizes so non-trivial functions of human intelligence as specifying of a problem to the well-formed one and solving of such well-formed problem.

Highly intelligent work does, however, not only the Solver of TULIPS but also its Analyser. Its goal is to extract from NL (Russian) input the descriptions of problem situations, user's goals, and own duties (represented in utterances both explicitly and implicitly). To extract implicit information the Analyser should take into consideration a context (local and global) of the current communication process. A local context of NL utterance is an information extracted from the previous input expressions. A global context is defined by models of "abstract" user, "external world" as a whole, particular users known for TULIPS, specific problem-domains (of user's activity), "general duties" of TULIPS, and concret heuristics and other tools related to a concret problem-domain or user.

These kinds of models must be introduced because of TULIPS is a widely-oriented system. The program is to solve different kindes of problems in different concret domains and to adapt to subset of Russian being used, specific features of problem-domains and users. This main feature of TULIPS - wide-orientation - determines all the other its features and, in particular, the approach to NL processing.

Approach to Natural Language Processing

Natural Language is a social level sign system that is used as a means of intercourse. It exists in a social group of speakers of a given language objectively - every member of this group should learn signs of NL_t rules of their combining - syntax, and the rules of interpretation and usage NL constructions - semantics, sigmatics, and pragmatics.

However, subset of NL learned by any speaker is specific by lexicon, meanings, and syntactic rules. The process of such a subset forming is a natural phenomenon depending on psychicial and social factors. At the same time speaker can't handle NL too arbitrarily. If he violates language norms appreciably (Humpty-Dumpty may be mentioned), he will not be understood by others. while language will not perform its fundamental social function - function of a means of intercourse.

Since a dependency of NL processing on a mentioned subjective factor is a real natural phenomenon, it must be taken into consideration in TULIPS model of language processing. Thus, on a conceptual level two basic concepts "meaning" and "sense" reflects, respectively, objective and subjective aspects of utterance content [1] •

The sense concept describes the aspects of the meaning that connects to an utterance current speaker (user) in a current real situation, whereas the meaning corresponds to standard objective Tor a social level) links of sign and:

1) frame and constituent sign structures - syntactic meaning;
2) objects and relations of reality - sigmstic meaning;
3) peychicial (reflective) equivalents of these objects and relations - semantic meaning;
4) activities, caused by speech acts and causing them - pragmatic meaning.

Thus, in a language communication any speaker relates to an utterance only a portion (sense) of objectively existing meaning - a portion that is the most relevant to a current act of speaking- While a hearer (no matter who a human or an artificial system) is to extract from utterance just this sense, to learn not only "what the speaker is saying" but also "what and whet for he intended to say".

NL processing of this kind is referenced here as understanding. To understand an utterance stands for to find thous aspects and portions of a meaning that are the most relevant to a current context situation, to find speaker's intention and to relate it to goals and activities of hearer.

It is obvious that a necessary feature of NL understanding (in the sense defined above) is using of data not only on a language but also on a current speaker, concret problem-domain and hearer's activities. That is why the vehicles of an understanding program are to be enough elaborate. They should be able to handle NL utterances, underlying structures, linguistics rules and to be adaptable. Non-trivial problem - in NL communication - is that of discovering an interpretation the most relevant (with respect to user and program) in a current intercourse act.

The Main Blocks of TULIPS

To discover such the most relevant interpretation of an utterance (to learn sense associated with it by user) TULIPS Analyzer - ANAL block - interactes with other blocks of the program. Thus, NL processing invokes the deductive mechanisms of TULIPS, problem-solving abilities, data represented in the model of external world domains and of user.

Interactions of ANAL with other blocks are two-way ones.ANAL block extracts en underlying conceptual content of an utterance and relates it to the models of current environment, while information represented in these models guides parsing and extracting of the conceptual content [2] . These interactions allow the program to insert NL processing in a total process of functioning.

The guiding of NL processing by context information is one of the fundamental features of the ANAL block. Another characteristic feature of this block processing concerns a method of enalysis. ANAL and its sub-blocks - MORP (morphological analysis) and SYNT (syntactic one) - handle corresponding fragment of an input utterance, being driven by data that describe (predicts) expected results of their processing.

Having recieved a prediction a block checks if an input satisfy it or not. If a prediction is satisfied, a result is returned to a parent block.* Otherwise a process of "failure-investigation" is initiated. It should be noted that the scheme of predictive analysis accepted - predictions besides syntactic categories describe semantic ones, references to user's goals and responsibilities of the program - allows to realize the principle of conceptually driven analysis.

Nevertheless a conceptual level orientation (orientation on the "supreme" aspects of a meaning - semantic and pragmatic ones) must not lead to ignoring of proper linguistic phenomena [3] That is why ANAL takes into account syntactic level characteristics - order of words, projectivity, grammatical agreement, and lexical relations (the powerful apparatus of "lexical functions" (4] is used). The vechicles of analysis implemented and detailed description of surface and deep levels of a natural language are the basis for discovering of subtle aspects of a meaning, that are hidden from both "non-intelligent" parser and "illiterate" AI system.

The TULIPS Solver - SOLV - is the most intelligent block of the program. Its main task is to plan a solution of a T-problem (i.e. a way to transform initial situation model - S - to a goal one - G) and to execute actions planed. The Solver can plan a solution in both directions. One of its main strategies is that of reduction of problems, basic general methods - analogy, induction, abstraction etc.

A problem-solving process is begining with analysis of S and G models, as a rule. At this stage the program makes a description of a problem - a pair (S,G) -

* As a rule, a prediction describes several possible results. Any block of the program specifies a prediction (by comparing it with input phrases) and returns selected concretized result.

more exact. A well-formed T-problem obtained - a triplet (S,F,G) - includes references to relevant to a current problem -domain means (P) - problem-solving methods. The second stage of solving (planning) returns a set of operations that transforms S to G. It is to be remarked that during the planning TULIPS can try to specify a T-problem again. The aim of repeted consideration of S and/or G is to discover some implicit properties of the situations - properties that can be found useful for a solving process. Sometimes TULIPS inquiere user for the aspects of real situations non-reflected in initial wording (and, hence, in S and G).

The main duty of SOLV is to solve a T-problem that corresponds to a user's problem, i.e. to plan and acomplish (on the world models) the situation transformations he haB told to the program. However, the Solver can show its own "activity" and generate T-problems represented in input utterances implicitly. The first kind of such a problem is a T-problem of sense discovering, the second - automatic extanding of linguistic knowledge to complete understanding of a new type phrase (NL adaptation).

The T-problems mentioned are generating during analysis of an input when ANAL appeals to the SOLV block. Common for both NL processing and non-linguiBtic domains is an implicit "super-problem" of teaching. The program's models of user and TULIPS itself contain assertions on a necessity of teaching. Hence the explicit user directions to memorize some new fact are optional.

The Synthesizer - SYNZ - generates Russian utterances adreseed for TULIPS's user (answers, requests etc.). SYNZ is the least elaborated and for the present time temporary block, which uses a few standard patterns to be filled by concret words and phrases.

Monitor-TULIPS - MTUL - is, in fact, a problem-oriented extension of LISP for BESM-6 computer. It implements flow of control among ANAL, SOLV, and SYNZ and embodies some other functions that are characteristic for new programming languages for AI research [5] . Moreover it accomplishes memory transformations (see below;.

Memory and Teaching

The problems of knowledge representation, memory structure, and mechanisms for memory handling are especially significant for TULIPS. In a concret Beance (the user fixed works with the program in fixed problem-domain) TULIPS utilizes only a small part of its knowledge. A lot of data items - CE (for a "conceptual element"), are irrelevant and some components of relevant ones may be smoothly ignored.

To take into account the relevance and validity of data-items the memory - structured set of CEs - is broken into several sub-fieldB and every CE is marked by a special relevation tag - RT. In accordance with the first criterion for the memory breaking up permanent memory (PSM) that contains unchangable in a qeance "absolute" data and operative memory (OSM) - for data valid during a seance - are introduced. The second criterion - relevance of data in a current domain - is a basis for introduction PSM zones. The total zone contains data that can be used in any seance. The particular zones describe specific properties of problem-domains and individual users.

Information read by TULIPS from particular zones guides NL analysis. Just this information is a vehicle to discover an utterance sense, to take into account both individual features of user speach and specific of language performance caused by problem-domain.

Memory Transformations

To use data from particular zones and data represented in relevant CEs from total zone TULIPS performs some operations on the memory. One of them is the activation that makes CEs accesible to the main TULIPS blocks. Active memory - a set of activated CEs - contains only relevant to a current seance CEs. This enables to speed up the necessary information retrieval.

An activation is the first stage of any seance. This process is initiated on receiving the first user sentence. A next utterances being processed, the new operations on active memory are performed. One of them is the reactivation - activation of CEs that were qualified firstly as irrelevant. Thie operation corresponds to such a creative human intelligence operation as using of irrelevant (a priory) information in a quite new domain.

Another operation - refinement - changes activated and already used in a current seance data-items. Refinement as opposed to activation (that rejects or accepts CEs as a whole) implies changing of CE's "body". It deletes irrelevant in a seance portions of CE and irrelevant references to other data-items.

The last type of memory transformation - adaptation - concerns changes of CEs grounding on a new knowledge on the world objects and relations, that the program learns in a seance.

Teaching

The adaptation process changes only OSM of the program - world model thai is valid during a current seance. For the new data-items to be available later on, all the changes should be memorized in the PSM. Such a memorizing - teaching - is, as a rule, accompanied by rather complicated actions: generalization, particular zone determination (zone for memorizing) etc. That is why a teaching is a task of a special MTUL sub-block that can invoke all the program abilities and starts on finishing a seance. The later TULIPS feature, as one can observe, makes the program teaching similar to human sleeping - period of sorting and memorizing during the day perceived information.

The adaptation and teaching processes may result in changing: 1) recomendatione on relevance - for unchanged data-items; 2) data-items as such- If the first kind of teaching takes a place, TULIPS changes either implicit recomendatione - localization in PSM zones, or explicit ones - RTB*. AS a new CE should be memorized in some zone of PSM and might be marked by RTs, the same transformations are, as a rule, performed, the changed CEs being memorized.

TULIPS may learn both NL processing and non-linguistic actions - in each of these spheres the same strategies and mechanisms are used. Thus, if the program obtains unexpected result it is equally valuable for any sphere of its activity, and "failure-investigator" may be invoked in this case both by SOLV (planned set of actions is impracticable, e.g.) and by ANAL (prediction is rejected;.

In the later case TULIPS searchs for a "culprit". As the program "belives" user language competency, it tries to adapt a lexicon and grammar. If meta-grammer - the model of morphological and syntactic rules (each rule has a description of its validity and changability) - allows, the appropriate changes are performed. Otherwise the program has to admit user as a culprit and to appeal to him with a request.

A seance is over, TULIPS sends new NL data into PSM. The concret zones determination is guided by meta-grammer: individual divergences from language norms are memorized in particular zone (for a current user), new words and rules may be send into any zone. This strategy allows the program (later on) to take into account features of a current user language-model•

Conclusion

For TULIPS - as for a widely-oriented AI system - both the sphere of problem-domains and language subset used have, in principle, no restrictions.That is why any a priory embodied knowledge can turn out insufficient. Thus, an ability to learn knowledge and to memorize them for future (teachability) proves to be a necessary feature of the program.

At present TULIPS sucessfuily teachs both Russian and the real situations model actions in the domaines:
1) planning actions (of user) in the simplest life situations;
2) solving of word-formed primary-school arithmetic problems (the initial version of TULIPS for this domain - the APRIL program - is described in [6]);
3) Russian-English translation (the new ad hoc version of SYNZ is used).

These domains having essentially different criteria of understanding, further experiments are of significant theoretical interest. To appraise the principles embodied, mechanisms used, and TULIPS universality more precisely the program should be taught to work in new more various domains.

References

1. Malkovsky M. 0. Analyser of Natural Language Understanding Program. Symbol InformationProcessing.v.2, Computing Center, USSR Academy of Sciences, Moscow, 1975.
2. Malkovsky M. 0. TULIPS - Natural Language Understanding Program. Proc. of ARSO-VIII. v.4, Lvov, USSR, 1974.
3. Malkovsky M. G. Natural Language Understanding Programs. Symbol InformationProcessing.v.1. Computing Center, USSR Academy of Sciences, Moscow, 1973.
4. Melchuck I. A. Essay on the Theory of Linguistic Models "sense<->text" "Nauka*, Moscow, 1974.
5. Bobrow D. and Raphael B. New Programming Languages for AI Research. ACM Computing Servevs. v.6, No.3, 1974.
6. Malkovsky M. G. APRIL - Program for Solving Word-Formed Arithmetic Problems. Algorithms and Algorithmic Languages, v.6. Computing Center. USSR Academy of Sciences, Moscow, 1973.

* RT markes not only CE as a whole but also its components. The later (internal) RTs are taken into account during the refinement process.