COMPUTER SUPPORT OF SEMANTIC TEXT ANALYSIS OF A TECHNICAL SPECIFICATION ON DESIGNING SOFTWARE

Alla Zaboleeva-Zotova, Yulia Orlova

Abstract: *The given work is devoted to development of the computer-aided system of semantic text analysis of a technical specification. The purpose of this work is to increase efficiency of software engineering based on automation of semantic text analysis of a technical specification. In work it is offered and investigated a technique of the text analysis of a technical specification is submitted, the expanded fuzzy attribute grammar of a technical specification, intended for formalization of limited Russian language is constructed with the purpose of analysis of offers of text of a technical specification, style features of the technical specification as class of documents are considered, recommendations on preparation of text of a technical specification for the automated processing are formulated. The computer-aided system of semantic text analysis of a technical specification is considered. This system consist of the following subsystems: preliminary text processing, the syntactic and semantic analysis and construction of software models, storage of documents and interface.*

Keywords: *natural language, semantic text analysis, technical specification.*

ACM Classification Keywords: I.2.7 Natural Language Processing

Conference: *The paper is selected from Second International Conference "Intelligent Information and Engineering Systems" INFOS 2009, Varna, Bulgaria, June-July 2009*

Introduction

Now designing software includes development of requirements or technical specification (TS), system or technical project (TP), programming or design engineering, trial operation, support and improvement. It is necessary to take into account interdependence of all basic parts of process of designing software from toolkit, technologies and the organizations of works. The majority of works to areas CAD systems are directed on creation and perfection of toolkit for automation of designing process. The significant contribution to development CAD systems have brought in V.I.Averchenkov, G.S.Altshuller, A.V.Andrejchikov, N.P.Buslenko, V.P.Bykov, B.S.Voinov, G.D.Volkova, V.Gasparsky, Дж. K.Dzhons, Дж. Dickson, M.F.Zaripov, V.A.Kamayev, K.V.Kumunzhiev, V.M.Kurejchik, P.M.Mazurkin, I.Muller, I.P.Norenkov, I.J.Petrova, A.I.Polovinkin, A.F.Pohilko, J.M.Solomentsev, F.Hanzen, P.Hill, A.Holl, etc.

Most known of the commercial software products used at designing of the software, basically are intended for visualization intermediate and end results of process of designing. Some of them allow to fully automate last design stages: generation of a code, creation of the accounting and accompanying documentation, etc. Thus the problem of automation of the initial stage of designing - formations and the analysis of the text of the technical project remains open. It is connected to extraordinary complexity of a problem of synthesis and the analysis of semantics of the technical text for which decision it is necessary to use methods of an artificial intellect, applied linguistics, psychology, etc. However, it is possible to come nearer to achievement of the given purpose, having allocated some small subtasks quite accessible to the decision by known methods of translation.

Proceeding from the aforesaid, it is possible to draw a conclusion, that the problem of creation of means for automation of process of designing is actual [1].

Ideas of a developed direction realization of the unified procedures of the designing equally answering to requirements of the expert - designer and requirements to technology to modelling of software products is main.

The purpose of this work is to increase efficiency of software engineering based on automation of semantic text analysis of a technical specification.

To achieve this purpose it is necessary to solve the following tasks:

1. To carry out the analysis of software engineering process and models of semantic text analysis;
2. To develop a technique of the text analysis of a technical specification;
3. To develop and investigate semantic model of the text of a technical specification;
4. To develop algorithmic maintenance of analysis of text of a technical specification and automatic construction of the software models;
5. To realize developed formalisms, a technique and algorithms as system of automation of the initial stage of designing software.

A Technique Of The Text Analysis Of A Technical Specification

In work it is offered and investigated a technique of the analysis of the text of a technical specification is submitted, the fuzzy attribute grammar of a technical specification, intended for formalization of limited Russian is constructed with the purpose of analysis of offers of text of a technical specification, style features of the technical specification as class of documents are considered, recommendations on preparation of text of a technical specification for the automated processing are formulated.

A technique of the analysis of the text of a technical specification consist of three stages: semantic text processing, creation of frame structure and creation of data flow diagrams of system described in the technical specification. (see Figure 1).

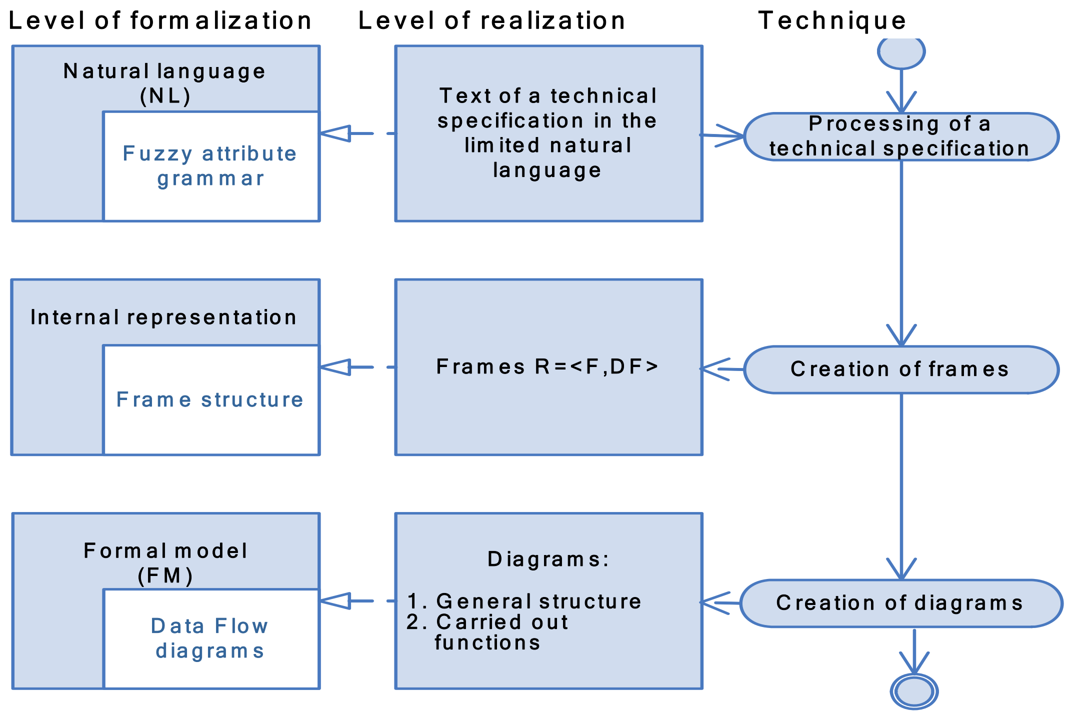

Figure 1: Technique Of The Text Analysis Of A Technical Specification

For realization of the first stage of a technique the semantic model of the text of a technical specification, including the requirements formulated as the document in the limited natural language has been developed; the second stage - the frame structure being internal representation of requirements; the third stage - model of software as the description of requirements in graphic language Data Flow Diagrams.

The semantic model of the text of a technical specification contains the developed expanded fuzzy attribute grammar above frame structure of the formal document "Technical specification" which allows to display contents TS most full.

The expanded fuzzy attribute grammar, necessary for the automated analysis of the text of a technical specification, is determined as:

AG = <N, T, P, S, B, F, A, D(A)>,

where N - final set of non-terminal symbols; T - not crossed with N set of terminal symbols; P - final set of rules; S - the allocated symbol from N, named an initial symbol; B - set of linguistic variables βk,i, corresponding to terminal symbols T (a variable i on k level); F - set of functions of a belonging fk,i, determining a degree of belonging mk,i linguistic variables βk,i; A - set of attributes, A = AsinUAsem, where Asin - syntactic attributes, Asem - semantic attributes; D (A) - final set of semantic actions. The fragment of grammar is submitted in table 1.

Table 1: Fragment of the developed fuzzy attribute grammar above frame structure of a technical specification

β_1	*<list of incoming data flows >*	<*incoming data flow name* > :: 'Name' <*incoming data flow description*> :: 'Contents' < *list of incoming data flows* >| ε
	<*incoming data flow description*>	*The text containing "entrance" or "entrance data "*:: 'Clause' <*incoming data flow*>::"Frame Data Flow=Creation ", "Input=Giving"
$\beta_{1,2}$	<*incoming data flow*>	[<*Number of data units*>]:: "Slot AMOUNT OF DATA = Giving" [<*Type of data*>]:: "Slot TYPE OF DATA = Giving" <*the Name of incoming data flow* >:: "Slot NAME OF INCOMING DATA FLOW = Giving"
β_2	<*function specification* >	<*function type* <*name of the functions liss*>:: 'Name'<*function description*>:: "Frame FUNCTION = Creation "*;* < *List of functions*> | ε
$\beta_{2,1}$	< *function type* >	«main» | «basic» | «additional»
$\beta_{2,2}$	<*function description* >	<*Name of function*>:: 'Name', "Slot NAME OF FUNCTION = Giving" <*List of incoming data flow*> <*List of outcoming data flow*>

Linguistic variables from set B = {βk,i}k,i used for the analysis of the text of a technical specification is described by the following five:

βk,i =<β, T(β), U, G, M>,

β - name of linguistic variable (basis for development, purpose of development, technical requirements to a program product, a stage and development cycles, etc.);

T(β) - language expressions. For linguistic variables of the top level they are the linguistic variables corresponding to terminals of the right part of a rule. For linguistic variables of the bottom level – fuzzy variables, that is expressions of a natural language.

U - Set of all probable values, $T(\beta) \subset U$;

G - rules of the morphological and syntactic description of language expressions which determine syntactic attributes Asin;

M - a semantic rule for linguistic variables which is induced by morphological and syntactic rules as the sense of a term in T is in part determined by its syntactic tree, and semantic attributes Asem.

Methods of representation connections between rules are broadcast on language of fuzzy mathematics. Thus connections are represented by fuzzy relations, predicates and rules, and sequence of transformations of these relations - as process of an fuzzy conclusion.

Linguistic variables of the top level are compound, that is include linguistic variables of the bottom level. Due to this it is possible to constract a tree of linguistic variables and to establish dependence between them.

Functions of an a belonging from set F = {fk,i}k,i linguistic variables {βk,i}k,i, are necessary for construction of an fuzzy conclusion. In particular, to each rule of grammar from set P function of a belonging fk,i is put in conformity. This dual system of substitutions is used for calculation of sense of a linguistic variable.

Actually grammar of a technical specification is used for splitting the initial text of the document into sections and processings of most important of them for our problem. It needs precise observance of structure of the document. Technical specification represents the structured text consisting of sequence of preset sections.

The frame structure of the technical specification is submitted as:

$$R = \left\langle N_R, \overline{F_R}, \overline{I_R}, \overline{O_R} \right\rangle$$

where N_R is a name of system, F_R is system functions vector, I_R is incoming data flows vector, O_R is outgoing data flows vector.

$$\overline{F}_R = \left\langle F_R^1, F_R^2, ..., F_R^k \right\rangle \text{, then } F_R^i = \left\langle N_F^i, \overline{I}_F^i, D_F^i, G_F^i, H_F^i, \overline{O}_F^i \right\rangle,$$

Where N_F^i - a name of function F_R^i, $\overline{I}_F^i$ - incoming data flows vector of F function, D_F^i - the name of the action which are carried out by function, G_F^i - subject of the function action, H_F^i - restrictions on function, $\overline{O}_F^i$ - a outgoing data flows vector of F function.

Let's denote the data flow by DF (Data Flow), then I_R, O_R, I_F, O_F are denoted by:

$$DF = \left\langle N_{DF}, D_{DF}, T_{DF}, C_{DF} \right\rangle$$

Where N_{DF} - data flow name, D_{DF} - data flow direction, T_{DF} - data type in flow, C_{DF} - data units per frame.

The model proposed is represented as a frame model with "a-kind-of" links (see Figure 2).

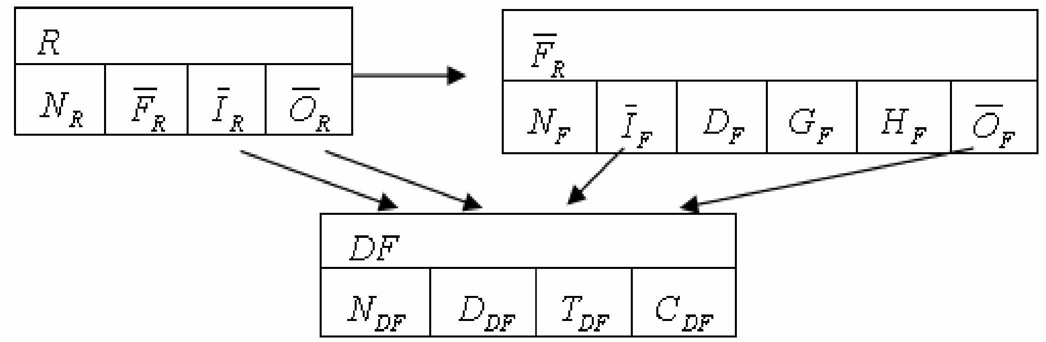

Figure 2: Frame network

Computer-Aided System Of Semantic Text Analysis Of A Technical Specification

The computer-aided system of semantic text analysis of a technical specification consists of the following subsystems: preliminary text processing, the syntactic and semantic analysis and construction of software models, storage of documents and interface (see Figure 3).

Figure 3: Architecture of computer-aided system of semantic text analysis of a technical specification

Preliminary text processing is necessary to share of a technical specification on separate lexemes. The incoming information of a subsystem is the text of a technical specification in the limited natural language, the target information - tables of sections, sentences and lexemes of a considered technical specification. Results can be submitted both as corresponding tables, and as a tree of sections.

Already after the first stage work not with the text of a technical specification, but with its parts submitted on sections is made. On a course of work of a technical specification shares all over again on more and more fine sections, then on separate sentences (with preservation of sections structure) and lexemes with the instruction of an accessory to sentences.

Preliminary text processing is carried out with use of final automatic device. During the work of final automatic device the symbols acting on its entrance, collect in the buffer. In the certain conditions of final automatic device record of the current contents of buffer in one of tables then the buffer is devastated is carried out. Work of automatic device proceeds up to achievement of a final condition.

After that the received tables act on an entrance of a subsystem of syntactic and semantic analysis. The semantic analysis of a text is made on the basis of the developed grammar of a text of technical specification.

Rules of top level serve for analysis of sections of top level. Rules for analysis of sections consist of two parts: the first part serves for analysis of a section name; the second part serves for analysis of a text contents in section. Symbols of the given grammar possess syntactic attributes. In attributes of non-terminal symbols names of frames or names of slots in which the information received during the further analysis should be placed are specified. Syntactic attributes of text can be in addition specified in attributes of terminal symbols. Comparison of words at analysis is made in view of their morphology. During analysis the syntactic and morphological analysis are made only in the event that there is such necessity that time of performance of semantic analysis is considerably reduced.

Let's consider a fragment of the developed attribute grammar submitted in a xml-format:

… <global-rule id="Section42" comment = "Section 4.2. Requirements to functional characteristics"> <rule><ruleref uri="#Section42Name"/><ruleref uri="#Section42x"/></rule></global-rule>

<global-rule id="Section42Name" sectionPart="Name" comment= "Heading of the unit 4.2."><rule><clause clauseType="UNCERTAIN"/><rule type="or"><words contains="Functions"/> <words contains= " functional characteristics "/> </rule></rule></global-rule>

<global-rule id="Section42x" frame= "FunctionFrame" frameSlot="Function" comment="Function"><rule> <ruleref uri="#Section42xName" /><ruleref uri="#Section42xContent" /> </rule></global-rule>

<global-rule id="Section42xContent" sectionPart="Content" comment="Inputs and outputs of function"><rule><ruleref uri= "#Section42xInputs" minOccurs="0"/><ruleref uri="#Section42xOutputs" minOccurs="0"/></rule></global-rule>

<global-rule id="Section42xInputs" comment="Inputs of function">

<rule><sentence/><clause/><rule type="or"><words contains="Inputs"/> <words contains="entrance data"/></rule><ruleref uri="#Input" maxOccurs="unbounded"/></rule></global-rule> …

The morphological and syntactic modules used in the program, are modules of the foreign developer. If in a rule of grammar there is a terminal having syntactic attribute the mechanism of syntactic analysis for current sentences is started [2].

After creation of a tree of analysis construction of frame description of a technical specification begins. For this purpose the information on frames and names of slots which contains in attributes of symbols of grammar is used.

The received frame structure contains the significant information about system: data about inputs and outputs of system, functions and restrictions. For each function inputs and outputs also are allocated. It allows to receive data flow diagrams of system which is described in a technical specification on the basis of frame structure.

The subsystem "Construction of data flow diagrams" carries out construction and ordering the column of data flows, and also creation the figures of data flow diagrams in Microsoft Office Visio.

For construction of data flows it is prospected of functions inputs conterminous to system inputs. Then functions on which all inputs data act, are located on the one level of diagram. Their inputs incorporate to system inputs. Further it is prospected functions which inputs coincide with outputs of functions received on the previous step. They are located on the following level, their inputs incorporate to outputs of the previous levels functions and with system inputs.

Work of algorithm proceeds until all functions will not be placed on the diagram. After that connection of function outputs with necessary system outputs is made.

The computer-aided system of semantic text analysis of a technical specification is developed on Microsoft .NET Framework 2.0 platform (language of development C#) using integrated development environment Visual Studio 2005.

Scientific Novelty

Scientific novelty consists in the following: a technique of text analysis of a technical specification at the initial stages of software engineering, including semantic model of text of a technical specification, transformation matter of text into the frame structure and construction of model of the software on its basis are developed.

Practical Value

Practical value of work is that as a result of development and introduction of a suggested technique quality of software engineering raises due to automation of routine work of the person on extraction of helpful information from standard documents and to displaying it as software models.

Conclusions and Future Work

Software designing differs from designing in other areas of a science and technics a little, therefore it is possible to expand results of the given work for application in other areas of human knowledge. Thus, opening prospects raise a urgency of the given work.

Bibliography

1. Kamsay, A. Computer-aided syntactic description of language systems/ A. Kamsay// Computational linguistics. An international handbook on computer-oriented language research and applications.- Boston: Walter de Giuyter, 1989.- P.204-218
2. Reyle, U. Natural language parsing and linguistic theories/ U. Reyle.- Berlin: Rohrer Dordrecht, 1998.- 625 p.
3. Tools Development For Computer Aided Software Engineering Based On Technical Specification`s Text Analysis / A.Zaboleeva-Zotova, Y.Orlova // Interactive Systems And Technologies: The Problems Of Human-Computer Interaction: Proc. of the Int. Conf., Ulyanovsk.

Authors' Information

Alla V. Zaboleeva-Zotova – PhD, professor; CAD department, Volgograd State Technical University, Lenin av., 28, Volgograd, Russia; e-mail: zabzot@vstu.ru

Yulia A. Orlova – PhD; CAD department, Volgograd State Technical University, Lenin av., 28, Volgograd, Russia; e-mail: yulia.orlova@gmail.com