COMPUTER-AIDED SYSTEM OF SEMANTIC TEXT ANALYSIS OF A TECHNICAL SPECIFICATION

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 the model of the analysis of the text of the technical project is submitted, the 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 project 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 consists 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 International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008

Introduction

Now designing of the software represents the labor-intensive process demanding of the user deep knowledge of a subject domain and skills in designing.

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 fully automating 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].

On CAD-department of the Volgograd state technical university questions of automation of designing of software products with use of natural - language support for a number of years are investigated.

The main ideas of the developed direction are:

- realization of the unified procedures of the designing equally answering to requirements of the expert
- design the requirements to technology to modeling of software products.

Designing of the software at the initial stages with use of a natural language is based on the following main principles:

1. Performance of all design procedures is modeled in language of internal representation of system. Internal representation is the unified model of designing of the software, based on methodology of the theory of systems and technologies of natural language processing.

2. A number of representations of the project is generated. Translation of a condition of the project into the certain language which is distinct from language of internal representation refers to as representation. Programming languages, natural languages or artificial formal languages of modeling of processes of designing can be attributed to such languages (UML, IDEF-diagrams, model of diagrams of streams of the data). Different representations reflect only separate aspects of the project.
3. Thus due to use of uniform internal model consistency of representations is provided.
4. The software of process of the designing, guaranteeing an opportunity of conducting the project on any of languages of representations is developed.
5. The basic language of representation of the project for the person - the customer and the designer - is the natural language. Dialogue between the customer and the designer is traditionally conducted in a natural language - language of human dialogue, but, as a rule, are entered new formal structures - diagrams, circuits, schedules. According to the developed concept, natural - language representation of the project supplements formal and serves as the tool facilitating understanding of process of designing.

As illustration of process of designing ON with use of the offered concept the diagram «to be», resulted on figure 1 serves.

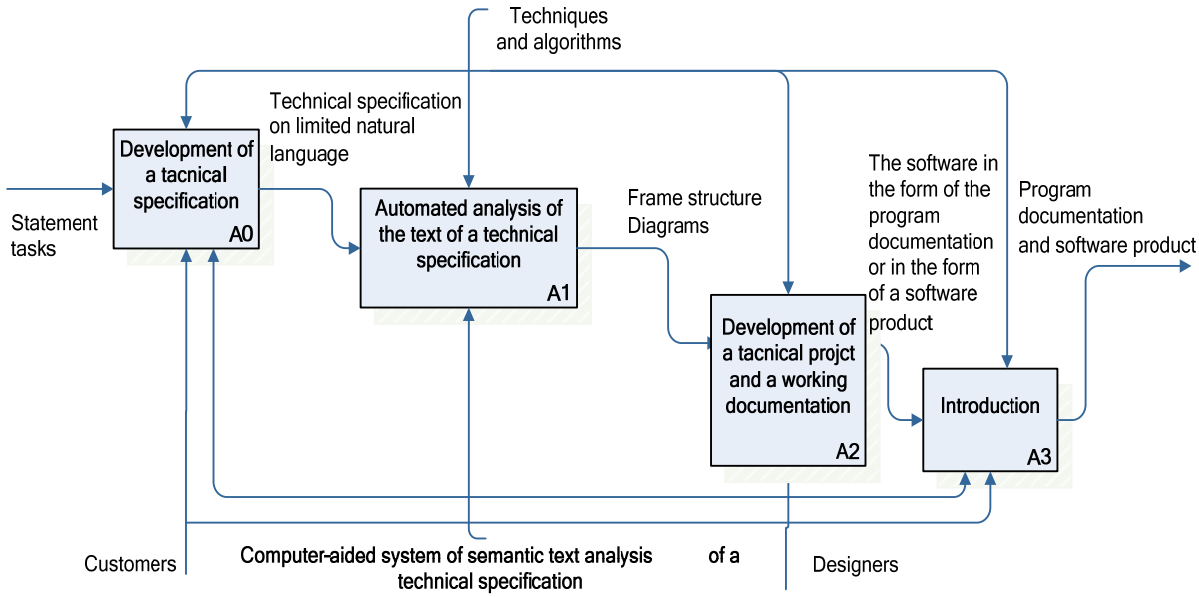

Figure 1: Diagram of process of designing «TO BE»

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.

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 and investigate model of the text analysis of a technical specification;
3. To develop a technique and analysis algorithms of text of a technical specification and construction of the software models;
4. To develop the computer-aided system of semantic text analysis of a technical specification;
5. To apply the system in software engineering process.

Model of the Text Analysis of a Technical Specification

In work it is offered and investigated the model of the analysis of the text of the technical project is submitted, the 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 project as class of documents are considered, recommendations on preparation of text of a technical specification for the automated processing are formulated.

Model input is the requirement specification written in the limited natural language, its output is a set of the data flow diagrams, describing the program system (see Figure 2).

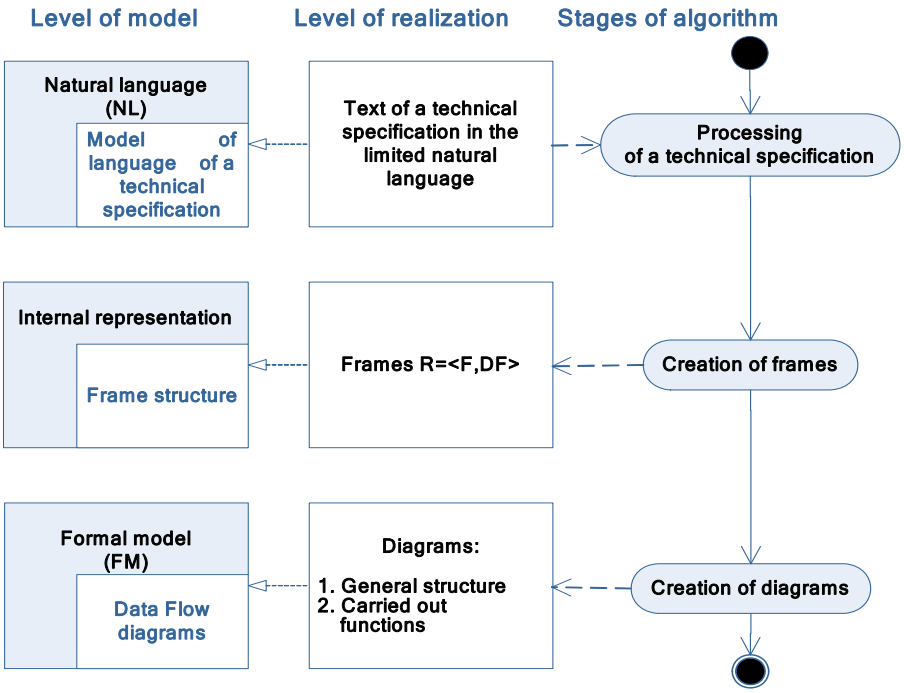

Figure 2: Model of the Text Analysis of a Technical Specification

The model consists of two levels:

1. Natural Language (NL), level of specification on the limited natural language.
2. Formal Model (FM), level of the description of limitations on the graphic Data Flow Diagrams language.

FM level includes the set of data flow diagrams:

1. Common system structure with the specification of incoming and outgoing data flows.
2. System functions and their incoming and outgoing data flows specifications.

According to the proposed model, the system can be considered as a black box. System behavior specification can be treated as a description of functions and data flows.

Structurally level NL can be divided into two parts: grammar of software specification on the limited natural language and frame model [3].

Grammar of software specification contains syntactic and semantic attributes (see Table1)

Table 1: Grammar of software specification

<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"**
<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"**
<function specification >	*< name of the functions liss >* :: **'Name'***< function description >*:: **"Frame FUNCTION = Creation "***; < List of functions> | ε**
< function type >	«main» | «basic» | «additional»
< function description >	*< Name of function>*:: **'Name', " Slot NAME OF FUNCTION = Giving "** *< List of incoming data flow > <List of out coming data flow>*

In connection with that the natural language is context-dependent for the description context-dependent grammars attributes are used. With their help it is possible to transfer the information from the left part of a generating rule in right and from the right part in left. Advantages attributive grammars that they can specify both context-free and context-dependent languages.

In rules of grammar there are syntactic and semantic attributes. For example, the syntactic attribute is underlined as follows: <*incoming data flow name* > :: 'Name', and semantic: < *function description* >:: "Frame FUNCTION = Creation" - the name of attribute and action.

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.

Level FM represents a set of Data flow diagrams: general structure of system with the indication of its entrance and target streams; the functions which are carried out by system with their entrance and target streams

Formally frame model can be described as R = <N_R, F_R, I_R, O_R>, 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 F_R= <N_F, I_F, D_F, G_F, H_F, O_F, where N_F is function name, I_F is incoming data flows vector of F function, D_F is function action, G_F is subject of the function action, , H_F is limitations and restrictions for F function, O_F is 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 = <N_{DF}, D_{DF}, T_{DF}, C_{DF}>, where N_{DF} is data flow name, D_{DF} is data flow direction, T_{DF} is data type in flow, C_{DF} is data units per frame.

The model proposed is represented as a frame model with "a-kind-of" links (see Figure 3).

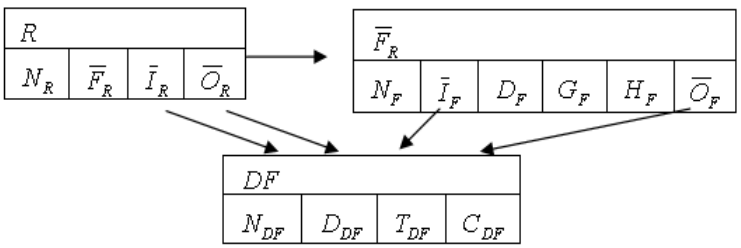

Figure 3: Frame network

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 4).

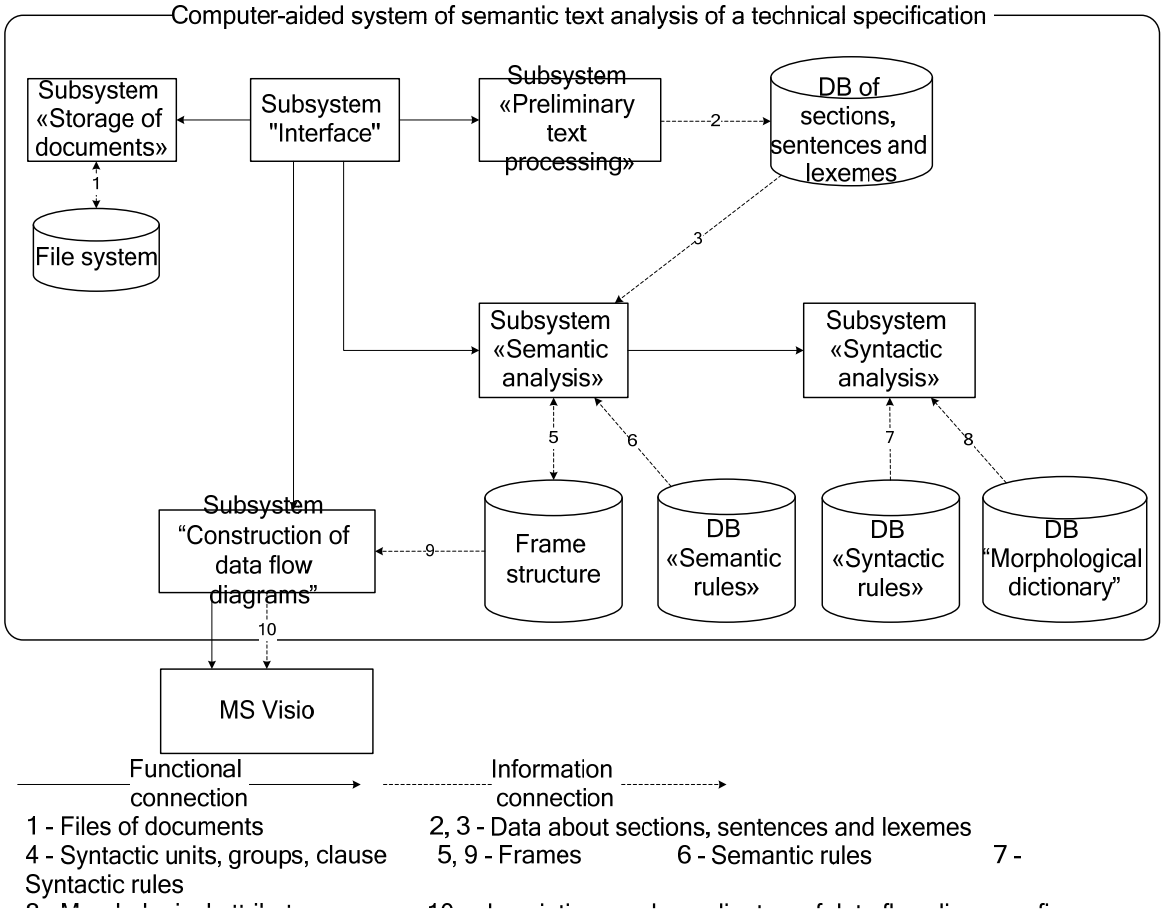

Figure 4: Architecture of computer-aided system of semantic text analysis of a technical specification

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 receiving 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: the model of text analysis of a technical specification at the initial stages of software engineering, including semantic model of text of a technical specification, the technique of transformation matter of text into the frame structure and construction of the 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 student; CAD department, Volgograd State Technical University, Lenin av., 28, Volgograd, Russia; e-mail: yulia.orlova@gmail.com