International Journal of Engineering & Technology, 7 (4.5) (2018) 489-491

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Configuration of the instrumental tools for generation of text representations of object of the curated database

Larisa Yu. Ismailova [1] *, Sergey V. Kosikov [2]

[1] National Research Nuclear University "MEPhI" (Moscow Engineering Physics Institute), Moscow, 115409 RF
[2] Institute for Contemporary Education "JurInfoR-MGU", Moscow, 119435 RF
**Corresponding author E-mail: kosikov.s.v@gmail.com*

Abstract

The paper considers the problem of configuring the instrumental tools (a workbench) for generating textual representations of objects in the curated database. A formal (formalized) representation of a fragment of a natural language sufficient to describe the content of database objects is proposed. Based on the matching of the expressive means of the proposed language, taking into account the general architecture of the comprehensive information system, the configuration of support tools is proposed. The configuration provides support to the procedures for manipulating formal descriptions, including the ability to process both formal and textual representation of database objects. The approach has been tested in the field of support to the implementation of best available technologies.

Keywords: *Applicative Grammar; Database Objects; Intentional Logic; Natural Language; Text Representation.*

1. Introduction

The curated databases [1] contain data selected by experts from many independent sources. The reliability and relevance of data is provided by using specialized confirmation processes, supported by the information system, containing the curated database. Such procedures are performed, as a rule, by professionals of a special profile - database supervisors, and can foresee both manual per- formance steps and automated procedures.

The data contained in such a database may have different representations. In the case of textual form, their processing requires the support of their representation in a formal (formalized) man- ner. The ability to verify the conformance of formalized and tex- tual representation of data is essential. For general type of data, the ability to generate a text description corresponding to the structure, format, and content of the data is essential.

It is possible to solve the indicated problems basing on the choice of a formal (formalized) representation of a fragment of the natural language sufficient to describe the content of database objects, and on the development of procedures for manipulating formal descriptions, including the possibility of processing both formal (formalized) and textual representations. The task of creating a workbench for supporting textual representations that maintain the curated database is of interest. This problem is considered in this paper.

The paper is structured as follows. In paragraph 2, the options for a formal (formalized) representation of fragments of the natural language are considered, the special attention being paid to means corresponding to the applicative approach. Paragraph 3 sets the problem to select a configuration of the workbench for processing formal (formalized) textual representations. Paragraph 4 gives brief characteristics to the considered languages of the representa- tion (or their fragments). Paragraph 5 proposes a solution of the problem in the form of a configuration of workbench that provides both the transition between the formal (formalized) and textual representations of objects and the transition from the representa- tion of the object in natural language to the representation of a logical type that is directly oriented to work with the database. The Conclusion gives a kind of summary.

2. Applicative textual representations

Formal (formalized) representations of sentences in the natural language were actively studied both within the framework of lin- guistics and at the junction of linguistics and informatics [2]. A generally accepted approach is the one that is based on the repre- sentation of the syntax of natural language sentences using trees. The approach has two varieties. The first variety - the so-called dependence trees - is based on the representation of the syntactic dependencies of words in the sentence. The second variety, simply called syntactic trees, allocates embedded (unfinished) phrases of different structure in the sentence and considers the methods of composition that allow receiving complete sentences from their integral parts. In the whole, the second approach, accepted in most researches [3], [4], is more convenient for computer processing.

It is necessary to especially highlight the applicative variety of the second approach, based on the so-called applicative grammar [5]. This approach represents the sentence of the natural language in the form of a structure, which is made up by the application of operators of higher order to the arguments. For example, in the sentence "the plant contaminates the river" the embedded phrase "contaminates the river" is considered as the application of opera- tor "contaminates" to the argument by the "river". The approach is interesting, in particular, by the fact that it can be coordinated with the applicative programming style (the well-known paradigm of functional programming).

It is also worth mentioning the paper [6] on the construction of the semantics of natural language, which proposes not only the means

for formalizing the syntax, but also considers the ways of representing the content of a sentence by means of formal logic. Within the framework of this direction, during the research on informatization of various areas of legal activity, the authors constructed a number of models [7, 8], providing, in particular, the representation of sentences of a limited fragment of natural language by means of intensional logic with the possibility of taking into account various ways of parameterization of the representation. The given work is focused on support of models of this type.

3. Task for configuring a workbench

The method of constructing a model of a fragment of the language that describes a domain, providing the representation of both the syntax and the content of the sentences of the fragment of the language under consideration, gives the basic possibilities for using it in the construction of the supporting tools for the informa- tion system. It, however, does not predetermine the configuration of the supporting tools for processing the considered representa- tions, which should include the following tools:

- development of object representations in the given syn- tax;
- transformation of representations from one syntax into another, including with the possibility to clarify the se- mantics in the process of transformation;
- checking the formalized properties of objects in a given representation (such as the property of typical correct- ness);
- Generation of representations that provide practical ma- nipulation of objects by the tools of the information sys- tem of the lower level.

The problem to determine the configuration of supporting tools for textual representations should be solved taking into account the need to harmonize these tools with the common architecture of the comprehensive information system. It seems, however, that it is possible to propose a general configuration of such tools that maintains the solution of problems typical for a rather wide range of tasks. It appears to be that such a configuration can be coordi- nated with the variants of typical architectures of information system.

4. Characteristics of the languages used

The solution of the studied problem involves the consideration of variants of representing the descriptions of objects in the form of sentences of a limited natural language, which in the future will be designated as NL, as well as various formalized languages. The main one among them is the language of syntactic trees, which will be designated as ST in what follows. The sentences of the language NL (ST, respectively) will be designated as e (respec- tively, t) and fix the fact that they belong to NL as e:NL (t:ST, respectively).

The fulfilled computational experiments make it possible to state that the configuration of the support tools depends rather weakly on the details of the formal definition of NL and ST. Therefore, their exact formal description is not included into the scope of the given paper. Nevertheless, it seems desirable to have a general (informal) definition of the abilities of languages, allowing their applicability for description of the objects of the information sys- tem.

The NL language is defined as an inductive class and contains:

1) Basic expressions (nouns and adjectives, verbs intransitive and transitive, pronouns, prepositions, etc.);
2) Application (usage) of quantitative expressions (“all”, “some”, “majority”, etc.) to expressions that allow such usage;
3) Definitional construction “x, which e(x)” or “x such, that e(x)”.
4) Application of operators to arguments (verbs to nouns, etc.);
5) logical facilities (expressions of the type “note”, “e1 and e2”, “e1 or e2”);
6) Means of substituting one expression for another;
7) Means of expression of time dependencies (“it was so that e” and “it will be so that e”).

The language ST is defined to a large extent in parallel to the language NL and for each expression of the language NL it contains either a base expression of the corresponding type or an operator that allows to construct a tree with subtrees corresponding to the basic expressions of the language NL. The main characteristic of the language ST as a whole is the ability to translate its sentences into the language NL. Let us designate the corresponding image as T: ST → NL.

An essential characteristic of a natural language is its ambiguity, which is the greatest challenge while modelling the natural language interaction. The proposed model takes into account this property of natural language on several levels.

Firstly, the one-to-one correspondence between the expressions ST and NL is not assumed. For some (but not for all) expressions e: NL there are such expressions t1: ST and t2: ST, that T (t1) = e and T(t2) = e. Thereby, to the expression e of the natural language can correspond two (or more) trees of its syntactic analysis. In this sense, the expression e is ambiguous, and this ambiguity is presented in the model.

Secondly, even having the unambiguous parsing of the expression e: NL in natural languages the situations are possible when various meanings can be given to it. The point is that the meaning of sentences of the natural language depends on the context. Contextual dependence can be accounted for in different ways [8]. The given paper characterizes the language NL as a limited natural language, since a simplifying assumption, related to it, is made; this assumption means that the context dependencies can be taken into ac- count with the help of subject-dependent axioms, limiting the class of language interpretations. With the indicated assumption, the meaning of the sentence of the natural language e:NL can be de- termined from the parse tree t:ST with the natural condition of the correspondence T(t) = e, taking into account the context depend- ences in the form of axioms of the limiting the class of interpreta- tions. In practice this condition is met for most dialects of business and technical prose (language of technical documentation, legal documents, etc.).

Along with the languages NL and ST, the formalized languages for representing the content of sentences NL will be also consid- ered. The language of intensional logic IL is the main among them is, and for this language a formal mapping of translation L:ST → IL can be defined (formal details are also omitted). A family of languages can also be defined as FR that is oriented to transfer the dependence of the values of ST expressions (and corresponding to them FR expressions) from the parameters.

5. Configuration of a workbench

The analysis of the problem of selecting the configuration of the workbench for generating textual representations of objects, taking into account the specific features of the languages for object description, allows proposing the following configuration of supporting tools for the processing of object representations.

The support is also assumed to be rendered to the converter of expressions ST into expressions NL. The reverse transformation due to its ambiguity is difficult to be implemented. Therefore, it is assumed to support the user’s interface tools to define expressions t: ST for the set ones e: NL and the following-up check of the relation T(t) = e.

The expressions IL of common type can include different intensional operators, corresponding to different ways of parameterization of language expressions. Therefore, it seems technically reasonable to define a family of frametype languages FR [7] that define the way of work with a particular type of parametrization. The corresponding methods of translation can be determined based on the general technique of abstraction. Common con- figuration of workbench is shown on Fig. 1.

The presented configuration of workbench has been tested when developing the workbench for legal information systems (exactly - while developing the systems in the field of environmental law for legal support to implement the best available technologies in the

Russian Federation). The testing showed the possibility of generating textual representations of information objects, meaningful for the user, and this makes it possible to state its applicability. At the same time the representation of phrases of a sufficiently deep level of nesting (about 4-7 depending on the type of operators used) appeared to be quite difficult for sensing. This one stimu- lates interest in the development of paraphrasing systems for de- scriptions of complicated applicative objects.

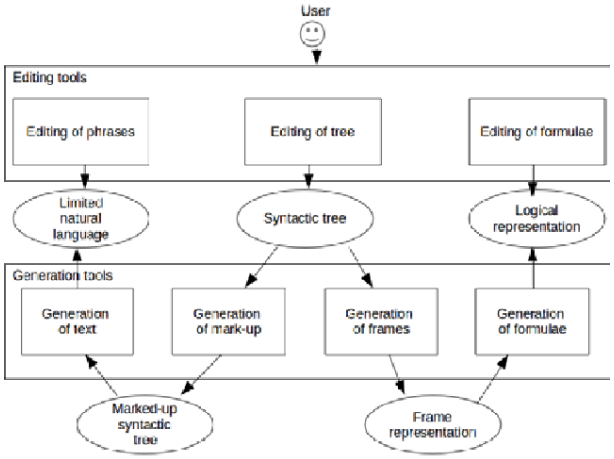

Fig. 1: Common Configuration of Workbench.

6. Conclusion

The paper considered the problem of providing a configuration of workbench for generating textual representations of the objects in the curated database, supporting:

- controlled transition from a formalized representation of an object to a grammatically correct textual representa- tion and vice versa;
- Generation of a logical expression, corresponding to a formalized representation of the syntax of the object.

The characteristics of a fragment of natural language that provide a description of the presentation set of objects of the curated databases have been proposed. For the selected characteristics, the configuration of workbench has been proposed that provides:

- development of a formal representation of the sentence, describing the object, and generation of the correspond- ing textual representation;
- generation of intermediate representations in frame-type languages that provide the representation of the content of the sentence with a controlled level of intensionality;
- Generation of the representation within the language of intensional logic, which gives the possibility of its use, in particular, when generating the database inquries.

The proposed configuration has been tested when generating textual representations of objects for legal profile tasks. The testing showed the principle applicability of the proposed approach and allowed to identify areas for further development of workbench.

Acknowledgement

The work has been supported by grants from the Russian Foundation for Basic Research (RFBR) 16-07-00909, 16-07-00912, 17-07-00893.

References

[1] Buneman P, Cheney J, Tan W-C, & Vansummeren S. “Curated da- tabases”. *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '08)*. (2008). ACM, New York, NY, USA, 1-12.

[2] Correia R, Mamede N, Baptista J & Eskenazi M. “Toward Automatic Classification of Metadiscourse”, *Advances in Natural Lan- guage Processing, 9th International Conference on NLP, PolTAL* 2014, Warsaw, Poland, September 17-19, (2014), pp. 262-269.

[3] Shenk R. *Processing of conceptual information*, M.: Energy, (1980), p. 360.

[4] Wolfengagen VE et al. “Evolutionary Domains for Varying Individuals”, *seventh Annual International Conference on Biologically In- spired Cognitive Architectures, BICA 2016, 16-19 July New York, USA*, Procedia Computer Science, Elsevier, Volume 88, (2016), Pages 347-352.

[5] Shaumyan SK, *Applicative grammar as semantic theory of natural languages*, Moscow: “Nauka”, (1974). – p. 204.

[6] Montague R. “Pragmatics”, *Klibanski, R. (ed.) Contemporary Phi- losophy*, Florence: La Nuova Italia Editrice, (1968). pp. 102—121.

[7] Ismailova LY, Kosikov SV, Wolfengagen VE, “Applicative Methods of Interpretation of Graphically Oriented Conceptual Information”, *7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016, 16-19 July New York, USA*, Procedia Computer Science, Elsevier, Volume 88, (2016), Pages 341-346.

[8] Ismailova LY, Kosikov SV, Wolfengagen VE, “A harmony and disharmony in mining of the migrating individuals”, *2016 Third In- ternational Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC)*, Moscow, (2016), pp. 52-57.