Semantics-Oriented Natural Language Processing

Volume 27

International Federation for Systems Research International Series on Systems Science and Engineering

IFSR was established "to stimulate all activities associated with the scientific study of systems and to coordinate such activities at international level." The aim of this series is to stimulate publication of high-quality monographs and textbooks on various topics of systems science and engineering. This series complements the Federation's other publications.

A Continuation Order Plan is available for this series. A continuation order will bring delivery of each new volume immediately upon publication. Volumes are billed only upon actual shipment. For further information please contact the publisher.

Volumes 1–6 were published by Pergamon Press.

Vladimir A. Fomichov

Semantics-Oriented Natural Language Processing

Mathematical Models and Algorithms

Springer

Vladimir A. Fomichov
Faculty of Business Informatics
Department of Innovations and Business
in the Sphere of Informational Technologies
State University – Higher School of Economics
Kirpichnaya Street 33, 105679 Moscow, Russia
vfomichov@hse.ru
vdrfom@aha.ru
v.fomichov@snhu.edu

Series Editor:

George J. Klir
Thomas J. Watson School of Engineering and Applied Sciences
Department of Systems Science and Industrial Engineering
Binghamton University
Binghamton, NY 13902
U.S.A

ISBN 978-0-387-72924-4 e-ISBN 978-0-387-72926-8
DOI 10.1007/978-0-387-72926-8
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009937251

Mathematics Subjects Classification (2000): 03B65, 68-XX, 93A30

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my wife, Olga Svyatoslavovna Fomichova

Preface

Gluecklich, die wissen, dass hinter allen
Sprachen das Unsaegliche steht.
Those are happy who know that behind
all languages there is something unsaid
Rainer Maria Rilke

This book shows in a new way that a solution to a fundamental problem from one scientific field can help to find the solutions to important problems emerged in several other fields of science and technology.

In modern science, the term "Natural Language" denotes the collection of all such languages that every language is used as a primary means of communication by people belonging to any country or any region. So Natural Language (NL) includes, in particular, the English, Russian, and German languages.

The applied computer systems processing natural language printed or written texts (NL-texts) or oral speech with respect to the fact that the words are associated with some meanings are called *semantics-oriented natural language processing systems (NLPSs)*.

On one hand, this book is a snapshot of the current stage of a research program started many years ago and called Integral Formal Semantics (IFS) of NL. The goal of this program has been to develop the formal models and methods helping to overcome the difficulties of logical character associated with the engineering of semantics-oriented NLPSs. The designers of such systems of arbitrary kinds will find in this book the formal means and algorithms being of great help in their work.

On the other hand, this book can become a source of new powerful formal tools for the specialists from several different communities interested in developing semantic informational technologies (or, shorter, semantic technologies), in particular, for the researchers developing

- the knowledge representation languages for the ontologies in the Semantic Web project and other fields;
- the formal languages and computer programs for building and analyzing the semantic annotations of Web sources and Web services;

- the formal means for semantic data integration in e-science and e-health;
- the advanced content representation languages in the field of multi agent systems;
- the general-purpose formal languages for electronic business communication allowing, in particular, for representing the content of negotiations conducted by computer intelligent agents (CIAs) in the field of e-commerce and for forming the contracts concluded by CIAs as the result of such negotiations.

During last 20 years, semantics-oriented NLPSs have become one of the main subclasses of applied intelligent systems (or, in other terms, of the computer systems with the elements of artificial intelligence).

Due to the stormy progress of the Internet, the end users in numerous countries have received technical access to NL-texts stored far away from their terminals. This has posed new demands to the designers of NLPSs. In this connection it should be underlined that several acute scientific – technical problems require the construction of computer systems being able to "understand" the meanings of arbitrary NL-texts pertaining to some fields of humans' professional activity. The collection of these problems, in particular, includes

- the extraction of information from textual sources for forming and updating knowledge bases of applied intelligent systems and the creation of a Semantic Web;
- the summarization of NL-texts stored on a certain Website or selected in accordance with certain criteria;
- conceptual information retrieval in textual databases on NL-requests of the end users;
- question answering based on the semantic-syntactic analysis of NL-texts being components of Webdocuments.

Semantics-oriented NLPSs are complex technical systems; their design is associated not only with programming but also with solving numerous questions of logical character. That is why this field of engineering, as the other fields of constructing complex technical systems, needs effective formal tools, first of all, the formal means being convenient both for describing semantic structure of arbitrary NL-texts pertaining to various fields of humans' professional activity and for representing knowledge about the world.

Systems Science has proposed a huge amount of mathematical models and methods that are useful for a broad spectrum of technical and social applications: from the design and control of airplanes, rockets, and ships to modeling chemical processes and production-sailing activity of the firms.

The principal purpose of this monograph is to open for Systems Science a new field of studies – the development of formal models and methods intended for helping the designers of semantics-oriented NLPSs to overcome numerous problems of logical character associated with the engineering of such systems.

This new field of studies can be called *Mathematical Linguocybernetics* (this term was introduced by the author in [66]).

Let's consider the informal definitions of several notions used below for describing the principal aspects of the scientific novelty of this book.

The term "semantics of Natural Language" will denote the collection of the regularities of conveying information by means of NL. *Discourses* (or narrative texts) are the finite sequences of the sentences in NL with the interrelated meanings.

If T is an expression in NL (a short word combination, a sentence, or a discourse), a *structured meaning of the expression T* is an informational structure being constructed by the brain of a person having command of the considered sublanguage of NL (Russian, English, or any other), and the construction of this structure is independent of the context of the expression *T*, that is, this informational structure is built on the basis of knowledge about only elementary meaningful lexical units and the rules of combining such units in the considered sublanguage of NL.

Let's agree that a *semantic representation (SR)* of an NL-expression *T* is a formal structure being either an image of a structured meaning of the considered NL-expression or being a reflection of the meaning (or content) of the given expression in a definite context – in a concrete situation of a dialogue, in the context of knowledge about the world, or in the context of the preceding part of the discourse.

Thus, an SR of an NL-expression *T* is such formal structure that its basic components are, in particular, the designations of the notions, concrete things, the sets of things, events, functions and relations, logical connectives, numbers and colors, and also the designations of the conceptual relationships between the meanings of the fragments of NL-texts or between the entities of the considered application domain.

Semantic representations of NL-texts may be, for instance, the strings and the marked oriented graphs (semantic sets).

An *algorithm of semantic-syntactic analysis* builds an SR of an NL-expression, proceeding from the knowledge about the morphology and syntax of the considered sublanguage of NL (English, Russian, etc.), from the information about the associations of lexical units with the units of conceptual level (or semantic level), and taking into account the knowledge about application domains. An SR of the text constructed by such an algorithm is interpreted by an applied computer system in accordance with its specialization, for instance, as a request to search an answer to a question, or a command to carry out an action by an autonomous intelligent robot, or as a piece of knowledge to be inscribed into the knowledge base, etc.

The scientific results stated in this monograph have been obtained by the author while fulfilling a research program started over 20 years ago. The choice of the direction of the studies was a reaction to almost complete lack in that time of mathematical means and methods that were convenient for designing semantics-oriented NLPSs.

The results of this monograph not only contribute to a movement forward but also mean a qualitative leap in the field of elaborating the formal means and methods of developing the algorithms of semantic-syntactic analysis of NL-texts. This qualitative leap is conditioned by the following main factors:

- The designers of NLPSs have received a system of the rules for constructing well-formed formulas (besides, a compact system, it consists of only ten main rules) allowing for (according to the hypothesis of the author) building semantic representations of arbitrary texts pertaining to numerous fields of humans' professional activity, i.e., SRs of the NL-texts on economy, medicine, law, technology,

politics, etc. This means that the effective procedures of constructing SRs of NL-texts and effective algorithms of processing SRs of NL-texts (with respect to the context of a dialogue or of a preceding part of discourse, taking into account the knowledge about application domains) can be used in various thematic domains, and it will be possible to expand the possibilities of these procedures in case of emerging new problems.

- A mathematical model of a broadly applicable linguistic database is constructed, i.e., a model of a database containing such information about the lexical units and their interrelations with the units of conceptual level that this information is sufficient for semantic-syntactic analysis of the sublanguages of natural language being interesting for a number of applications.
- A complex and useful, strongly structured algorithm of semantic-syntactic analysis of NL-texts is elaborated that is described not by means of any programming system but completely with the help of a proposed system of formal notions, this makes the algorithm independent of program implementation and application domain.
- A possible structure of several mathematical models of the new kinds is proposed with the aim of opening for Systems Science a new field of studies high significance for Computer Science.

Informational technologies implemented in semantics-oriented NLPSs belong to the class of *Semantic Informational Technologies* (or, shorter, *Semantic Technologies*). This term was born only several years ago as a consequence of the emergence of the Semantic Web project, the use of ontologies in this project and many other projects, the elaboration of Content Representation Languages as the components of Agent Communication Languages in the field of Multiagent Systems, and of the studies on formal means for representing the records of negotiations and the contracts in the field of Electronic Commerce (E-commerce).

One of the precious features of this monograph is that the elaborated powerful formal means of describing structured meanings of NL-texts provide a broadly applicable and flexible formal framework for the development of Semantic Technologies as a whole.

Content of the Book

The monograph contains two parts. Part 1, consisting of Chaps. 1, 2, 3, 4, 5, and 6, will be of interest to a broad circle of the designers of Semantic Informational Technologies. Part 2 (Chaps. 7, 8, 9, 10, and 11) is intended for the designers of Semantics-Oriented Natural Language Processing Systems.

Chapter 1 grounds the necessity of enriching the inventory of formal means, models, and methods intended for designing semantics-oriented NLPSs. Special attention is paid to showing the necessity of creating the formal means being convenient for describing structured meanings of arbitrary sentences and discourses pertaining to various fields of humans' professional activity. The context of Cognitive

Linguistics for elaborating an appropriate approach to solving this problem is set forth. The possible structure of mathematical models of several new kinds for Systems Science is outlined.

The basic philosophical principles, history, and current composition of an original approach to formalizing semantics of NL are stated in Chap. 2; this approach elaborated by the author of this book is called Integral Formal Semantics (IFS) of Natural Language.

In Chap. 3, an original mathematical model describing a system of primary units of conceptual level used by applied intelligent systems is constructed and studied. The model defines a new class of formal objects called conceptual bases.

In Chap. 4, based on the definition of a conceptual basis, a mathematical model of a system of ten partial operations on structured meanings (SMs) of NL-texts is constructed. The essence of the model is as follows: using primary conceptual units as "blocks," we are able to build with the help of these ten partial operations the structured meanings of the texts – sentences and discourses – from a very rich sublanguage of NL (including articles, textbooks, the records of commercial negotiations, etc.) and to represent arbitrary pieces of knowledge about the world.

The model determines a new class of formal languages called standard knowledge languages (SK-languages) and can be interpreted as a formal metagrammar of a new kind. A mathematical study of the properties of SK-languages is carried out. In particular, the unambiguity of the syntactical analysis of the expressions of SK-languages is proved.

The purpose of Chap. 5 is to study the expressive possibilities of SK-languages. The advantages of the theory of SK-languages in comparison, in particular, with Discourse Representation Theory, Theory of Conceptual Graphs, Episodic Logic, and Database Semantics of Natural Language are analyzed.

Chap. 6 shows a broad spectrum of the possibilities to use the theory of SK-languages for solving a number of acute problems of Computer Science and Web Science. The possibilities of using SK-languages for (a) building semantic annotations of informational sources and of Web services; (b) constructing high-level conceptual descriptions of visual images; (c) semantic data integration in e-science, e-health, and other e-fields are indicated.

The definition of the class of SK-languages can also be used for the elaboration of formal languages intended for representing the contents of messages sent by computer intelligent agents (CIAs). It is also shown that the theory of SK-languages opens new prospects of building formal representations of contracts and records of commercial negotiations carried out by CIAs.

The broad expressive power of SK-languages demonstrated in Chaps. 4, 5, and 6 provides the possibility to propose in the final part of Chap. 6 a new, theoretically possible strategy of transforming evolutionarily, step by step, the existing Web into a Semantic Web of a new generation.

In Chap. 7, a broadly applicable mathematical model of *linguistic database* is constructed, that is, a model of a collection of semantic-syntactic data associated with primary lexical units and used by the algorithms of semantic-syntactic analysis for building semantic representations of natural language texts.

Chapter 8 sets forth a new method of transforming an NL-text (a statement, a command, or a question) into its semantic representation (SR). One of the new ideas of this method is the use of a special intermediary form of representing the results of semantic-syntactic analysis of an NL-text. This form is called a Matrix Semantic-Syntactic Representation of the introduced text. The constructed SR of an NL-text is an expression of a certain SK-language, or a K-representation of the considered NL-text. A pure syntactic representation of an analyzed text isn't used: the proposed method is oriented at directly finding the conceptual relations between the fragments of an NL-text.

Chapters 9 and 10 together describe an original, complex, and strongly structured algorithm of semantic-syntactic analysis of NL-texts; it is called the algorithm *SemSynt*1. Chapter 9 sets forth an algorithm of constructing a Matrix Semantic-Syntactic Representation of a natural language text; this algorithm is called *BuildMatr*1. The algorithm *BuildMatr*1 is multilingual: the input texts may belong to the sublanguages of English, German, and Russian languages (a Latin transcription of Russian texts is considered).

Chapter 10 describes an algorithm *BuildSem*1 of assembling a K-representation of an NL-text, proceeding from its matrix semantic-syntactic representation. The final algorithm *SemSynt*1 is defined as the composition of the algorithms *BuildMatr*1 and *BuildSem*1.

The content of Chaps. 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 can be interpreted as the principal part of *the theory of K-representations (knowledge representations)* – a new, powerful, and flexible framework for the development of semantic technologies.

The final Chap. 11 discusses two computer applications of the obtained theoretical results. The first one is a computer intelligent agent for fulfilling a semantic classification of e-mail messages. The second one is an experimental Russian-language interface implemented in the Web programming system PHP on the basis of the algorithm *SemSynt*1, it transforms NL-descriptions of knowledge pieces (in particular, definitions of concepts) first into the K-representations and then into the expressions of the ontology mark-up language OWL.

Moscow, Russia *Vladimir Fomichov*
December 2008

Acknowledgements

I am highly grateful to Distinguished Professor George J. Klir for the kind invitation to prepare this book and numerous constructive advices.

Many thanks to the anonymous referees of my monograph proposal and of the submitted manuscript.

The love, patience, and faith of my wife, Dr. Olga Svyatoslavovna Fomichova, and my mother-in-law, Dr. Liudmila Dmitrievna Udalova, very much helped me to obtain the results reflected in this book.

Two persons provided an outstanding help in the preparation of the electronic version of this book.

My son Dmitry, an alumnus of the Faculty of Cybernetics and Computational Mathematics of the Lomonosov Moscow State University, spent a lot of effort for translating from Russian into English the main content of Chaps. 9 and 10, in particular, for following up the tradition of stepped representation of the algorithms while preparing a LaTex file with the description of the algorithm of semantic-syntactic analysis *SemSynt*1.

I greatly appreciate the generous LaTex consulting provided by my friend, Senior Researcher in Software Development, Alexandr Mikhailovich Artyomov. Besides, special thanks to Alexandr for transforming my hand-drawn pictures into the electronic files with the figures presented in this book.

I appreciate the creative approach to the preparation of diploma work by the alumni of the Moscow State Institute of Electronics and Mathematics (Technical University), Faculty of Applied Mathematics, Inga V. Lyustig and Sergey P. Liksyutin. The applied computer systems based on the theory of K-representations and designed by Inga and Sergey are described in the final chapter of this book.

The communication with the Springer US Editors, Ms. Vaishali Damle and Ms. Marcia Bunda, during various stages of the work on the book highly stimulated the preparation of the manuscript.

Contents

Acronyms

B	conceptual basis
BuildMatr1	algorithm of constructing an MSSR of an NL-text
BuildSem1	algorithm of semantic assembly
Cb	marked-up conceptual basis
Ct, Ct(B)	concept-object system
Det	morphological determinant, M-determinant
F, F(B)	the set of functional symbols
Frp	dictionary of prepositional semantic-syntactic frames
Gen, Gen(B)	generality relation on a set of sorts
Int, Int(B)	the set of intensional quantifiers
Lec	a set of basic lexical units
Lingb	linguistic basis
Ls(B)	SK-language, standard knowledge language
Lsdic	lexico-semantic dictionary
Morphbs	morphological basis
Matr	matrix semantic-syntactic represenation
Morphsp	morphological space
P, P(B)	distinguished sort “a meaning of proposition”
R, R(B)	the set of relational symbols
Rc	classifying representation of an NL-text
ref, ref(B)	referential quantifier
Rm	morphological representation of an NL-text
Rqs	dictionary of the role interrogative word combinations
S, S(B)	sort system
Spmorph	morphological space
S, St(B)	the set of sorts
SemSynt1	algorithm of semantic-syntactic analysis
Tform	text-forming system
Tol, Tol(B)	tolerance relation on a set of sorts
X(B)	primary informational universe
V(B)	the set of variables
Vfr	dictionary of verbal – prepositional semantic-syntactic frames
W	a set of words