
SEMANTIC COMPUTING

Edited by

Phillip C.-Y. Sheu

Heather Yu

C. V. Ramamoorthy

Arvind K. Joshi

Lotfi A. Zadeh



IEEE PRESS



A JOHN WILEY & SONS, INC., PUBLICATION

SEMANTIC COMPUTING

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

IEEE Press Editorial Board
Lajos Hanzo, *Editor in Chief*

R. Abari
J. Anderson
F. Canavero
T. G. Croda

M. El-Hawary
B. M. Hammerli
M. Lanzerotti
O. Malik

S. Nahavandi
W. Reeve
T. Samad
G. Zobrist

Kenneth Moore, *Director of IEEE Book and Information Services (BIS)*

SEMANTIC COMPUTING

Edited by

Phillip C.-Y. Sheu

Heather Yu

C. V. Ramamoorthy

Arvind K. Joshi

Lotfi A. Zadeh



IEEE PRESS



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2010 by the Institute of Electrical and Electronics Engineers, Inc.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey. All rights reserved.

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Semantic computing / Phillip Sheu ... [et al.].

p. cm.

ISBN 978-0-470-46495-3 (cloth)

1. Semantic computing. I. Sheu, Phillip C.-Y.

QA76.5913.S46 2010

006.7—dc22

2009033995

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE	ix
CONTRIBUTORS	xi
1. Semantic Computing	1
<i>Phillip C.-Y. Sheu</i>	
PART I SEMANTIC ANALYSIS	11
2. What Computers Need to Know About Verbs	13
<i>Susan Windisch Brown and Martha Palmer</i>	
3. Content-Based Retrieval of Videos	33
<i>Min Chen, Shu-Ching Chen, and Mei-Ling Shyu</i>	
4. A Semantic and Detection-Based Approach to Speech and Language Processing	49
<i>Li Deng, Kuansan Wang, and Rodrigo Capobianco Guido</i>	
5. Link Analysis in Web Mining: Techniques and Applications	69
<i>Prasanna Desikan, Colin DeLong, and Jaideep Srivastava</i>	
6. Conversational Thread Extraction and Topic Detection in Text-Based Chat	87
<i>Paige Adams and Craig Martel</i>	
7. Speaker Recognition and Diarization	115
<i>Gerald Friedland and David van Leeuwen</i>	
8. Machine Learning Methods for Ontology Mining	131
<i>Nicola Fanizzi, Claudia D'Amato, and Floriana Esposito</i>	

9. Visual Ontology Construction and Concept Detection for Multimedia Indexing and Retrieval	155
<i>Lyndon Kennedy and Shih-Fu Chang</i>	
10. Process Mining and Description	181
<i>Qi Wang and Phillip C.-Y. Sheu</i>	
PART II SEMANTIC LANGUAGES AND INTEGRATION	199
11. Semantic Integration: The Hawkeye Approach	201
<i>Jeff Hefflin and Zhengxiang Pan</i>	
12. Semantics of Software Modeling	229
<i>Fernando Silva Parreiras, Steffen Staab, Jürgen Ebert, Jeff Z. Pan, Krzysztof Miksa, Harald Kühn, Srdjan Zivkovic, Stefano Tinella, Uwe Assmann, and Jakob Henriksson</i>	
13. Semantic-Driven Component-Based Automated Code Synthesis	249
<i>Jicheng Fu, Farokh B. Bastani, and I.-Ling Yen</i>	
14. Semantic Web Services	285
<i>Shu Wang, Qi Wang, Haitao Gong, and Phillip C.-Y. Sheu</i>	
PART III SEMANTIC APPLICATIONS	301
15. Knowledge Work	303
<i>Aljoscha Burchardt, Iryna Gurevych, and Max Mühlhäuser</i>	
16. Searching Videos in Visual Semantic Spaces	323
<i>Eric Zavesky, Zhu Liu, Dave Gibbon, and Behzad Shahraray</i>	
17. Social and Expert Research in Online Communities	341
<i>Florian Metze, Christian Bauckhage, and Tansu Alpcan</i>	
18. Phase Coherence in Conceptual Spaces for Conversational Agents	357
<i>Giorgio Vassallo, Giovanni Pilato, Agnese Augello, and Salvatore Gaglio</i>	
19. Role of Semantics in Question Answering	373
<i>Dan Moldovan, Marta Tatu, and Christine Clark</i>	
20. Developing and Using a National Cross-Domain Semantic Web Infrastructure	421
<i>Eero Hyvönen</i>	

21. Semantic Analysis for Multimedia Security Application	439
<i>Fadi Almasalha, Faisal Bashir, Ashfaq Khokhar, Farrukh Khan, Hammad Haseeb, and Arif Ghafoor</i>	
 PART IV SEMANTIC INTERFACE	 465
22. Semantic Languages for Software Engineering	467
<i>Shu Wang, Fei Xie, and Phillip C.-Y. Sheu</i>	
23. From Semantic Objects to Structured Natural Language	489
<i>Phillip C.-Y. Sheu, Atsushi Kitazawa, Chihiro Ishii, Kenichi Kaneko, and Fei Xie</i>	
 INDEX	 507

Semantic computing is emerging quickly as an interdisciplinary area that is full of promises. Tremendous amount of research has been done in the past and there is a lot more to be done in the future.

For the past three years, we have tried to give semantic computing a more precise definition; the following is a summary of our attempts, published at various times by the Institute for Semantic Computing (homepage, <http://www.isc-home.org>).

2007: “Semantic Computing addresses the computing technologies (e.g., artificial intelligence, natural language, software engineering, data and knowledge engineering, computer systems, signal processing, etc.), and their interactions, that may be used to extract or process the Content and Semantics of multimedia, texts, services as well as structured data.”

2008: “The field of Semantic Computing (SC) brings together those disciplines concerned with connecting the (often vaguely formulated) intentions of humans with computational content. This connection can go both ways: retrieving, using and manipulating existing content according to user’s goals (‘do what the user means’); and creating, rearranging, and managing content that matches the author’s intentions (‘do what the author means’).”

2008: “Computing facilitated by using the descriptions of content.”

2009: “Semantic Computing is defined as a field that addresses the derivation and matching of the semantics of computational content and that of naturally expressed user intentions to help retrieve, manage, manipulate or even create the content, where ‘content’ may be anything including video, audio, text, process, service, hardware, network, community, etc. It brings together those disciplines concerned with connecting the (often vaguely formulated) intentions of humans with computational content. This connection can go both ways: retrieving, using and manipulating existing content according to user’s goals (‘do what the user means’); and creating, rearranging, and managing content that matches the author’s intentions (‘do what the author means’).”

Until now, its “definition” is still evolving; perhaps it will not have a precise definition for good. What we know for sure is that “semantics” may play a major role in future computing.

The intention of this book is not trying to define what exactly “semantic computing” is. Rather, it is a humble effort to bring together researchers from several areas, including natural language processing, software engineering, multimedia, semantic Web, and semantic services and share with the reader what we think may contribute to a better definition of semantic computing.

Many subjects are missing in the book. One obvious reason is that we are limited by our knowledge about the broad coverage of semantic computing. Another reason may be that some subjects are already well known in the literature (e.g., data mining, keyword-based search engines). Yet another reason may be that such knowledge does not exist yet (e.g., semantic computer architecture?)

The book is divided into four parts, based on a preliminary architecture of a semantic computing system: *semantic analysis*, *semantic integration*, *semantic application*, and *semantic programming and interface*.

The book may be considered an attempt to bring together researchers with different backgrounds based on an assumption that we may have worked “together” on a common problem: finding and utilizing the “semantics” to better access or manage “content” of some sort. Hopefully a synergy may be created so we can all work even more closely to solve or address some grand challenges of computing we are facing today.

We would like to thank Ms. Mary Mann and Lisa Van Horn of Wiley for their outstanding editorial support. We thank all authors for contributing their original work to the book and sharing their wisdom with us.

PHILLIP C.-Y. SHEU
HEATHER YU
C. V. RAMAMOORTHY
ARAVIND KOSHI
LOTFI. A. ZADEH

CONTRIBUTORS

Paige Adams, Naval Post Graduate School, Monterey, California

Agnese Augello, Institute of High Performance Computing and Networking, Palermo, Italy

Fadi Almasalha, College of Engineering, Purdue University, West Lafayette, Indiana

Tansu Alpcan, Carnegie Mellon University, Pittsburgh, Pennsylvania

Uwe Assmann, University of Koblenz-Landau, Koblenz, Germany

Faisal Bashir, College of Engineering, Purdue University, West Lafayette, Indiana

Farokh B. Bastani, Department of Computer Science, University of Texas, Dallas, Texas

Christian Bauckhage, Carnegie Mellon University, Pittsburgh, Pennsylvania

Susan Windisch Brown, University of Colorado, Boulder, Colorado

Aljoscha Burchardt, Saarland University, Saarbrücken, Germany

Shih-Fu Chang, Electrical Engineering, Columbia University, New York, New York

Min Chen, Florida International University, Miami, Florida

Shu-Ching Chen, Florida International University, Miami, Florida

Christine Clark, Lymba Corporation, Richardson, Texas

Claudia D'Amato, Università degli Studi di Bari, Bari, Italy

Colin DeLong, Naval Postgraduate School, Monterey, California

Li Deng, Microsoft Research Corporation, Redmond, Washington

Prasanna Desikan, Naval Postgraduate School, Monterey, California

Jürgen Ebert, University of Koblenz-Landau, Koblenz, Germany

Floriana Esposito, Università degli Studi di Bari, Bari, Italy

Nicola Fanizzi, Università degli Studi di Bari, Bari, Italy

Gerald Friedland, ICSI, University of California, Berkeley, California

Jicheng Fu, Department of Computer Science, University of Texas, Dallas, Texas

Salvatore Gaglio, Institute of High Performance Computing and Networking, Palermo, Italy

Arif Ghafoor, College of Engineering, Purdue University, West Lafayette, Indiana

Dave Gibbon, AT&T Research, Middletown, New Jersey

Haitao Gong, University of California, Irvine, California

Rodrigo Capobianco Guido, Institute of Physics at Sao Carlos, University of Sao Paulo, Sao Paulo, Brazil

Iryna Gurevych, Saarland University, Saarbrücken, Germany

Hammad Haseeb, College of Engineering, Purdue University, West Lafayette, Indiana

Jeff Hefflin, Lehigh University, Bethlehem, Pennsylvania

Jakob Henriksson, University of Koblenz-Landau, Koblenz, Germany

Eero Hyvönen, Helsinki University of Technology, Helsinki, Finland

Chihiro Ishii, NEC Soft, Koto, Japan

Kenichi Kaneko, NEC Soft, Koto, Japan

Lyndon Kennedy, Columbia University, New York, New York

Ashfaq Khokhar, College of Engineering, Purdue University, West Lafayette, Indiana

Atsushi Kitazawa, NEC Soft, Koto, Japan

Harald Kühn, University of Koblenz-Landau, Koblenz, Germany

Zhu Liu, AT&T Research, Middletown, New Jersey

Craig Martel, Naval Postgraduate School, Monterey, California

Florian Metze, Carnegie Mellon University, Pittsburgh, Pennsylvania

Krzysztof Miksa, University of Koblenz-Landau, Koblenz, Germany

Dan Moldovan, Lymba Corporation, Richardson, Texas

Max Mühlhäuser, Saarland University, Saarbrücken, Germany

Martha Palmer, University of Colorado, Boulder, Colorado

Jeff Z. Pan, University of Koblenz-Landau, Koblenz, Germany

Zhengxiang Pan, Lehigh University, Bethlehem, Pennsylvania

Fernando Silva Parreiras, University of Koblenz-Landau, Koblenz, Germany

Giovanni Pilato, Institute of High Performance Computing and Networking, Palermo, Italy

Behzad Shahraray, AT&T Research, Middletown, New Jersey

Min Shen, Florida International University, Miami, Florida

Phillip C.-Y. Sheu, University of California, Irvine, California

Mei-Ling Shyu, Florida International University, Miami, Florida

Jaideep Srivastava, Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota

Steffen Staab, University of Koblenz-Landau, Koblenz, Germany

Marta Tatu, Lymba Corporation, Richardson, Texas

Stefano Tinella, University of Koblenz-Landau, Koblenz, Germany

David van Leeuwen, ICSI, University of California, Berkeley, California

Giorgio Vassallo, Institute of High Performance Computing and Networking, Palermo, Italy

Kuansan Wang, Microsoft Research Corporation, Redmond, Washington

Qi Wang, University of California, Irvine, California

Shu Wang, University of California, Irvine, California

Fei Xei, University of California, Irvine, California

I.-Ling Yen, Department of Computer Science, University of Texas, Dallas, Texas

Eric Zavesky, AT&T Research, Middletown, New Jersey

Srdjan Zivkovic, University of Koblenz-Landau, Koblenz, Germany

Semantic Computing

PHILLIP C.-Y. SHEU

We define semantic computing as a field that addresses the derivation and matching of the semantics of computational content and that of naturally expressed user intentions to help retrieve, manage, manipulate, or even create the content, where “content” may be anything including video, audio, text, process, service, hardware, network, community, and so on. It brings together those disciplines concerned with connecting the (often vaguely formulated) intentions of humans with computational content. This connection can go both ways: retrieving, using, and manipulating existing content according to user’s goals (“do what the user means”) and creating, rearranging, and managing content that matches the author’s intentions (“do what the author means”).

1.1 CONNECTIONS BETWEEN CONTENT AND INTENTIONS

The connection between content and the user can be made via (1) semantic analysis, which analyzes content with the goal of converting it to a description (semantics); (2) semantic integration, which integrates content and semantics from multiple sources; (3) semantic services, which utilize content and semantics to solve problems; and (4) service integration, which integrates different kinds of service to provide more powerful services; and (5) semantic interface, which attempts to interpret naturally expressed user intentions (Fig. 1.1). The reverse connection converts descriptions of user intentions to create content of various sorts via techniques of analysis and synthesis.

Note that as most of the information is sent and received through a network, security is needed at multiple levels including data-level [1], communication level [1], database level [2], application level [3] and system (community) level

A part of this chapter is revised from P. C.-Y. Sheu, Editorial Preface, *International Journal of Semantic Computing*, 1.1:1–9, 2007.

Semantic Computing, edited by Sheu, Yu, Ramamoorthy, Joshi, and Zadeh
Copyright © 2010 the Institute of Electrical and Electronics Engineers, Inc.

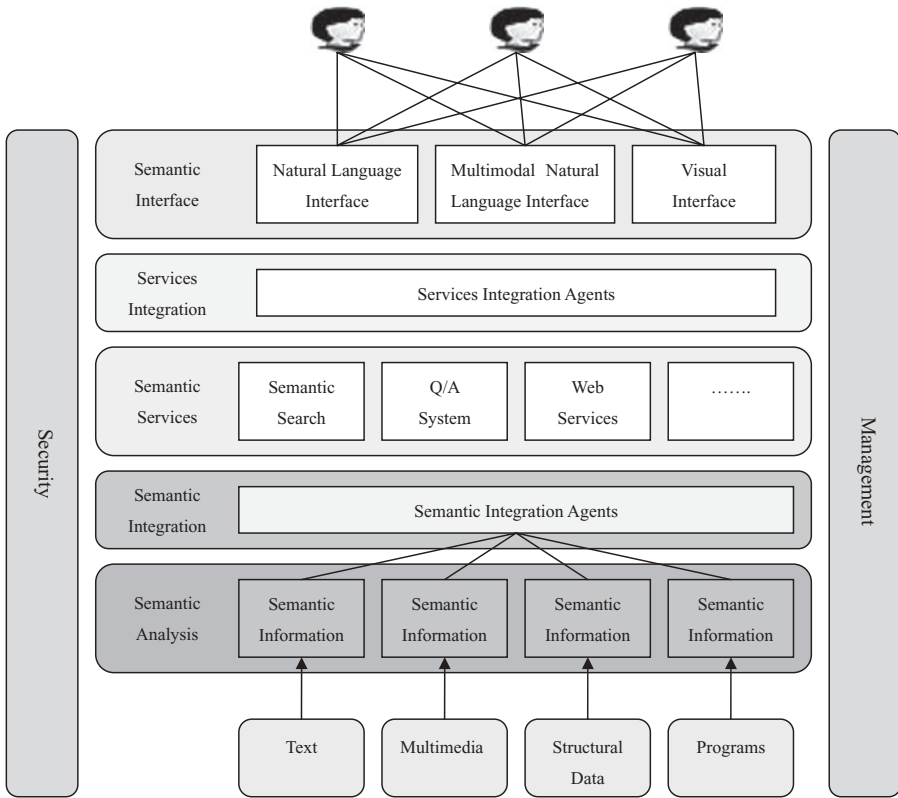


Figure 1.1 Architecture of semantic computing.

[4]. The flows of information are controlled both horizontally and vertically to assure desirable properties including QoS (quality of services) [5, 6] and integrity.

1. *Semantic Analysis*—analyzes and converts signals such as pixels and words (content) to meanings (semantics).
2. *Semantic Integration*—integrates the content and semantics from different sources with a unified model; it also includes languages and methodologies needed for developing semantic applications.
3. *Semantic Services*—utilize the content and semantics to solve problems, and some applications may be made available to other applications as services.
4. *Service Integration*—integrates different services to provide more powerful service.
5. *Semantic Interface*—allows the user intentions to be described in a natural form.

1.2 SEMANTIC ANALYSIS

“Semantic analysis is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the text as a whole, to their language-independent meanings, removing features specific to particular linguistic and cultural contexts, to the extent that such a project is possible. The elements of idiom and figurative speech, being cultural, must also be converted into relatively invariant meanings.”¹ Semantic analysis is the foundation of semantic computing; it provides the information resource for semantic integration and semantic services.

The research areas related to semantic analysis include but are not limited to:

1. Natural language understanding and processing
2. Understanding and processing of texts and multimodal content
3. Understanding of texts, images, videos, and audios
4. Speech recognition
5. Web mining
6. Data mining
7. Process mining

Semantic analysis may be the most developed part among the five layers of semantic computing, but it still has a lot of limitations. Most research on semantic analysis has focused on multimedia data analysis [7–9], text analysis (including shallow semantic parsing [10], latent semantic analysis [11, 12], as well as probability latent semantic analysis [13]), structural data analysis [14], and web analysis [15]. Early attempts at semantic analysis addressed complex problems such as semantic understanding, knowledge representation, and reasoning, and some progress has been reported on the understanding of certain domain-specific stories [16, 17]. The success, however, largely depends on domain-specific knowledge. A more robust approach is yet to be developed.

The output of semantic analysis is a description of content. To be useful, the description has to be machine processable. Several languages have been proposed to support such descriptions, including keywords, ontology, Moving Pictures Experts Group (MPEG), and others. In the case that automatic semantic analysis is difficult, descriptions may be generated manually in the form of annotations.

1.3 SEMANTIC INTEGRATION

Semantic integration considers the descriptions derived from the semantic analysis layer that is presented in different formats and integrates such information before it can be used. Existing work on semantic integration includes:

¹<http://www.wikipedia.org>.

- *Database Schema Integration* [18, 19] Database schemas may have different structures. Schema integration aims at unifying the matching elements. Various data sources are integrated into a data warehouse.
- *Data Exchange* [20, 21] To enable data exchange, applications need to convert messages between the formats required by different trading partners.
- *Ontology Integration (or Merging)* [22–24] Given two distinct and independently developed ontologies, it produces a fragment that captures the intersection of the original ontologies. This area is similar to that of schema integration but is more difficult in nature due to the rich and complex knowledge representation structures found in ontologies.
- *Ontology Mapping* [25, 26] Ontology mapping could provide a common layer from which several ontologies could be accessed by multiple applications. Ontology mapping is different from integration and merging because it does not try to combine several ontologies into a single, complete one.
- *Semantic Conflict Resolution* [30] This is needed to ensure semantic interoperability among heterogeneous information sources.

1.4 SEMANTIC SERVICES

A major goal of semantic computing is providing more powerful computing services for all kinds of users. Semantic services have been developed in several forms:

- *Web search*, including automatic question answering (Q/A) [32, 33] and information retrieval (e.g., Google² and Windows Live³). Both have attracted significant amount of attention in the past and at present.
- *Multimedia databases*, with a primary focus on content-based retrieval [34].
- *Domain-specific applications*, designed to support interoperable machine-to-machine interactions over a network for specific applications.⁴

1.5 SERVICES INTEGRATION

Although semantic services are useful for different kinds of users, sometime they are limited or insufficient for applications requiring several services working together.

A prerequisite for a set of services to collaborate is their ability to understand the mental model, often described in the form of ontology, of each other

²<http://www.google.com>.

³<http://www.live.com>.

⁴<http://www.w3.org/TR/ws-arch/>.

and communicate with each other. Mapping between ontologies is a major area of interest where automated and scalable solutions are also sought due to the vast number of services. Service integration [27–29, 31] provides the inter-operation methods between different services involved in practical scenarios.

On the other hand, for service integration, a significant gap still exists between specialists and nonspecialists, or among specialists focusing on different aspects of a problem. Traditional web services provide a protocol UDDI to perform resources description, discovery and integration. However, this protocol can only be used by experienced specialists. Automatic composition of services hence is strongly desired.

1.6 SEMANTIC INTERFACE

To achieve the goal of providing more powerful computing services to all kinds of users, a portable and friendly user interface is required. This is especially important when cell phones become more capable. Standard graphical user interface (GUI) techniques such as browsing, menu trees, and online help may be far less appealing for the next-generation applications. Therefore new standards of interface, such as natural language interface, multimodal interface, and visual interface, are becoming increasingly important.

A *natural language interface* allows people to interact using a form of a human language, such as English, as opposed to a computer language, a command line interface, or a GUI. Natural language interfaces may be designed for understanding either written or spoken texts. Attempts have been made to replace command lines and database queries with some form of natural language queries and to use some natural language syntax for programming languages. The most common problem of a true natural language interface is ambiguity: The same sentence may have multiple interpretations. Another problem is that users often assume that computers can reason as a human being and has the full knowledge as a human being.⁵

A *multimodal natural language interface* combines natural language input with other forms of input such as gesture. There are several strong reasons for creating an interface that allows voice to be combined with gesture as the input [35]:

- Expression is easy.
- Voice and gesture complement each other and when used together create an interface more powerful than either modality alone.
- Combining speech and gesture may improve the accuracy of recognition and reduce the length of speech, resulting in faster task completion compared to using speech alone.
- Users work more efficiently by using speech and gesture together.

⁵http://www.usabilityfirst.com/glossary/term_755.txt.

In addition to understanding user intentions, a semantic interface should be able to present the result produced by a semantic application in a form that can be easily understood by the user. In online analytical processing (OLAP), for example, it is important to provide a reporting tool on top of the server. Research on visualization allows the user to effectively visualize complex systems of information/data; it is particularly useful for decision making, training, simulation, virtual reality, augmented reality, and wearable computing applications [36].

Semantic programming is essentially another aspect of semantic interface: It allows users to express, in a natural way, their intentions when creating content which may be a program, a video, a process, a document, or others.

1.7 SUMMARY

Some areas of semantic computing have appeared as isolated pieces in various fields such as computational linguistics, artificial intelligence, multimedia, software engineering, database, and services computing. As shown in Figure 1.2, the field of semantic computing *glues these pieces together into an integrated theme and addresses their synergetic interactions*. For example, it addresses how retrieval of multimedia content may be facilitated by natural language annotations, how embedded texts may be extracted from images, how software may be derived from requirements described in natural language, how security can be added based on contexts, how Web search can be accomplished effectively with a cell phone, and so on.

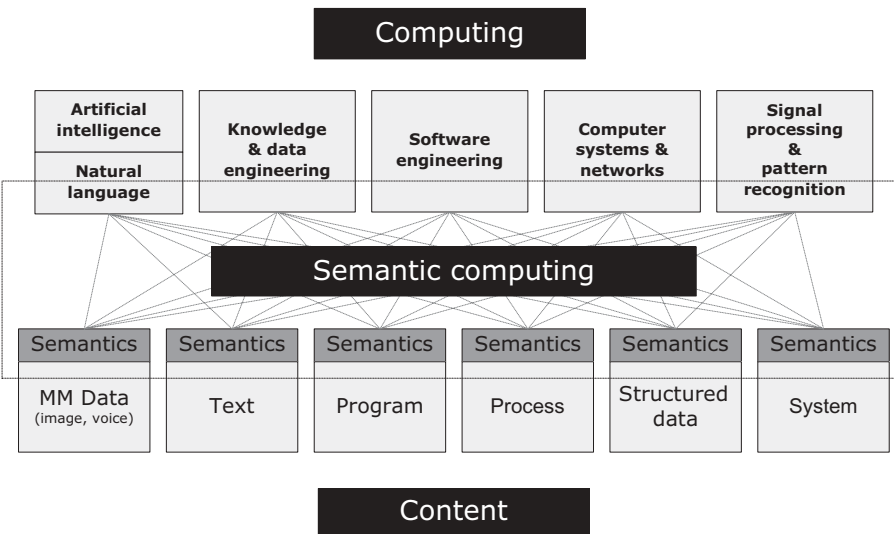


Figure 1.2 Technical coverage of semantic computing.

This may be the first book ever that attempts to introduce semantic computing as an integrated discipline. While researchers in the past have focused on their individual fields, considering semantic computing as an integrated discipline has the advantage that people may share their approaches in solving common problems. More importantly, more applications require the integration of different types of content and their corresponding tools to address complex requests from the user.

REFERENCES

1. W. Stallings, *Cryptography and Network Security: Principles and Practices*, Prentice-Hall, Englewood Cliffs, NJ, 1998.
2. S. Castano, M. G. Fugini, G. Martella, and P. Samarati, *Database Security*, ACM Press/Addison-Wesley, New York, 1995.
3. J.D. Meier et al., *Improving Web Application Security, Threats and Countermeasures*, Microsoft Corporation, Portland, OR, 2003.
4. R. Anderson, *Security Engineering*, Wiley, New York, 2001.
5. M. Ni and X. Xiao, Internet QoS: A big picture, *IEEE Network*, 13(2):8–18, 1999.
6. C. Aurrecoechea, A. Cambell, and L. Hauw, A survey of QoS architectures, *Multimedia Systems*, 6(3):138–151, Springer Berlin, Heidelberg, 1998.
7. A. Ekin, A. M. Tekalp, and R. Mehrotra, Integrated semantic-syntactic video modeling for search and browsing, *IEEE Trans. Multimedia*, 6(6):839–851, 2004.
8. S. Bloehdorn, N. Simou, V. Tzouvaras, K. Petridis, S. Handschuh, Y. Avrithis, I. Kompatsiaris, S. Staab, and M. G. Strintzis, Knowledge representation for semantic multimedia content analysis and reasoning, in *Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, 25–26, Paola Hobson, Ebroul Izquierdo, Ioannis Kompatsiaris and Noel E. O'Connor (Eds.): *Knowledge-Based Media Analysis for Self-Adaptive and Agile Multi-Media*, QMUL, London, 2004.
9. S. Bloehdorn, K. Petridis, C. Saatho, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab, and M. G. Strintzis, Semantic annotation of images and videos for multimedia analysis, in *Proceedings of the Second European Semantic Web Conference*, 592–607, Gómez-Pérez, Asuncion; Euzenat, Jerome (Eds.): *The Semantic Web: Research and Applications*, Springer Berlin, Heidelberg, Heraklion, Crete, Greece, 2005.
10. D. Lin, Automatic retrieval and clustering of similar words, paper presented at 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, *COLINGACL'98*, Montreal, 1998.
11. T. K. Landauer, P. W. Foltz, and D. Laham, Introduction to latent semantic analysis, *Discourse Process.*, 25:259–284, 1998.
12. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, Indexing by latent semantic analysis, *J. Am. Soc. Inform. Sc.*, 41(6):391–407, 1990.
13. C. Ding, A similarity based probability model for latent semantic indexing, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research*

- and *Development in Information Retrieval*, 58–65, Fredric Gey, Marti Hearst, Richard Tong (Eds.), ACM Press, Berkeley, 1999.
14. S. Nestrov, S. Abiteboul, and R. Motwani, Extracting schema from semistructured data, in *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, 295–306, Ashutosh Tiwary and Michael Franklin (Eds.), ACM Press, Seattle, 1998.
 15. S. Mehmet, S. Akhil, V. Machiraju, and F. Casati, *Semantic Analysis of E-Business Operations*, *Journal of Network and Systems Management*, 11(1), 13–37, Springer, New York, 2003.
 16. S. Soderland, Learning information extraction rules for semi-structured and free text, *Machine Learning*, 34(1–3):233–272, 1999.
 17. T. K. Landauer and S. T. Dumais, A solution to Plato’s problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge, *Psychol. Rev.*, 104:211–140, 1997.
 18. C. Batini, M. Lenzerini, and X. Navathe, A comparative analysis of methodologies for database schema integration, *ACM Comput. Surv.*, 18(4):323–364, 1986.
 19. J. Hammer, and D. McLeod, An approach to resolving semantic heterogeneity in a federation of autonomous, heterogeneous database systems, *Int. J. Intell. Cooperative Inform. Syst.*, 2:51–83, 1993.
 20. R. Fagin, P. G. Kolaitis, R. J. Miller, and R. Popa, Data exchange: Semantics and query answering, in *Proceedings of the International Conference on Database Theory (ICDT)*, 207–224, Diego Calvanese, Maurizio Lenzerini and Rajeev Motwani (Eds.): *Database Theory—ICDT 2003*, Springer Berlin, Siena, Italy, 2003.
 21. J. Madhavan and A. Y. Halevy, Composing mappings among data sources, in *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 29: 572–583, Johann Christoph Freytag, Peter C. Lockemann, Serge Abiteboul, Michael J. Carey, Patricia G. Selinger and Andreas Heuer (Eds.), VLDB Endowment, Berlin, Germany, 2003.
 22. D. Calvanese, G. De Giacomo, and M. Lenzerini, A framework for ontology integration, in *Proc. of 2001 Int. Semantic Web Working Symposium (SWWS)*, 303–316, Isabel F. Cruz, Stefan Decker, Jérôme Euzenat, Deborah L. McGuinness (Eds.): *The Emerging Semantic Web, Selected Papers from the First Semantic Web Working Symposium*, IOS Press, Stanford University, Palo Alto, CA, 2001.
 23. A. Doan and A. Halevy, Semantic integration research in the database community: A brief survey, *AI Mag.*, 26(1):83–94, 2005.
 24. N. Noy, Semantic integration: A survey of ontology-based approaches, *SIGMOD Record*, Volume 33, No. 4, 65–70, 2004.
 25. Y. Kalfoglou and M. Schorlemmer, Ontology mapping: The state of the art, *Knowledge Eng. Rev.*, 18(1):1–31, 2003.
 26. N. F. Noy and M. A. Musen, Evaluating ontology-mapping tools: Requirements and experience, paper presented at the Workshop on Evaluation of Ontology Tools at EKAW’02 (EON2002), Siguenza, Spain, 2002.
 27. L. Levine, B. C. Meyers, E. Morris, P. R. H. Place, and D. Plakosh, System of systems interoperability: Final report, SEI TR-004, Carnegie Mellon Software Engineering Institute, Pittsburgh, PA, 2004.
 28. J. Park and S. Ram, Information systems interoperability: What lies beneath? *ACM Trans. Inform. Syst. (TOIS)*, 22(4):595–632, 2004.

29. A. M. Ouksel and A. Sheth, Semantic interoperability in global information systems, *ACM SIGMOD Record*, 28(1):5–12, 1999.
30. S. Ram and J. Park, Semantic conflict resolution ontology (SCROL): An ontology for detecting and resolving data and schema-level semantic conflicts, *IEEE Trans. Knowledge Data Eng.*, 16(2):189–202, 2004.
31. G. Jiang et al., IXO Seedling Project technical report dynamic integration of distributed semantic services, Thayer School of Engineering, Dartmouth College, Hanover, NH, 2002.
32. J. Heflin and J. Hendler, Searching the Web with SHOE, in *Artificial Intelligence for Web Search. Papers from the AAAI Workshop*, WS-00-01, Bollacker K (Ed.), AAAI Press, Menlo Park, CA, 2000, pp 35–40.
33. E. Brill, S. Dumais, and M. Banko, An analysis of the AskMSR question-answering system, in *Proc. Empirical Methods in Natural Language Processing Conference*, 10:257–264, Jan Hajic, Yuji Matsumoto (Eds.), Association for Computational Linguistics, Philadelphia, 2002.
34. V. S. Subrahmanian, *Principles of Multimedia Database Systems*, Morgan Kaufmann, San Francisco, 1998.
35. O. Stock, Natural language in multimodal human-computer interface, *IEEE Expert*, 9(2):40–44, 1994.
36. S. N. Murphy, V. Gainer, and H. C. Chueh, A visual interface designed for novice users to find research patient cohorts in a large biomedical database, *AMIA Annu. Symp. Proc.*, Volume 2003, 489–493, 2003.
37. Jim Jacobs and Alexander Linden, Technical Report, T-17-5338, Gartner, Inc., http://www.gartner.com/DisplayDocument?doc_cd=109295.

PART I

SEMANTIC ANALYSIS

What Computers Need to Know About Verbs

SUSAN WINDISCH BROWN and MARTHA PALMER

2.1 INTRODUCTION

Verbs are generally considered the core of a sentence because they provide complex information about events or situations and about the relationships between the other items in the sentence. In fact, the particular verb used influences these items' selection, expression, and connection to each other so strongly that they are often called "arguments" of the verb. Given the crucial role of the verb in correctly processing the meaning of a sentence, an ideal computational lexicon would provide a rich representation of a verb's semantic and syntactic patterns. In this chapter we will try to define the criteria for such a lexicon, taking into consideration evidence from psycholinguistics, the fluid nature of language, and the needs of various language processing applications. We will review existing lexical resources, including WordNet, VerbNet, PropBank, and FrameNet, and evaluate how well they match these criteria.

Exactly how useful these lexicons will be for current statistical natural language processing (NLP) systems remains an open question. In the glory days of rule-based systems, lexicons played a central role and were often considered the heart of the system. Transfer-based and interlingua-based machine translation systems still rely heavily on rich bilingual or multilingual lexicons that not only pair source-language verbs with their closest translation in the target language but also map each argument of the source-language verb to its closest equivalent in the target language, even when this requires transforming a grammatical role like subject or object into a prepositional phrase or adjunct. However, more recently, natural language processing has striven to escape the bounds of domain-specific applications through

the use of statistical techniques which offer the promise of accurate processing over a wide range of text corpora. The majority of these approaches are supervised, shifting the effort from the hand crafting of grammars and lexicons to the hand annotation of vast amounts of training data. Linguistics expertise goes into developing the guidelines and example sentences that are used by the annotators. The systems are expected to induce appropriate lexical representations for sentences from sets of individual annotated examples.

If all of the lexical knowledge is represented explicitly or implicitly in the annotation, one might ask: What purpose can be served by additional free-standing lexical resources? In addition to the key roles that WordNet, FrameNet, VerbNet, and PropBank play in providing guidance to annotators, as detailed below, there are three other major potential contributions which have yet to be fully realized. The first has to do with actual attainment of the goal of broad coverage, rather than its current illusion. The NLP field is only too familiar with the degradation in performance that occurs when syntactic parsers, for example, those that have been trained on the *Wall Street Journal* (WSJ), are tested on different corpora. This was highlighted recently when CoNLL 2005 semantic role labeling (SRL) systems trained on the WSJ were tested on the Brown Corpus [1]. The Charniak POSagger degrades by 5%, and the Charniak parser *F* score by 8%, from 88.25 to 80.84. For the 19 systems participating in the SRL evaluation, there was in general a 10% performance decrease from WSJ to Brown. Unseen vocabulary is a major factor in the performance degradation on new corpora. At the moment the systems cannot extend their coverage successfully to lexical items that do not appear in the training data. This could change through the use of empirical connections between class-based lexical resources and training data examples. Systems could do a better job of handling unseen vocabulary items that occur in these resources in classes that are well represented in the training data by extrapolating from individual examples to classes and then back to other class members. In this way lexicons that provide information about syntactic and semantic properties that are shared between lexical items could provide important class-based backoff guidance.

In addition, these same lexicons could form the basis of empirically based bilingual lexicons derived from parallel proposition banks (as defined in Section 2.4.5) which could improve the range and accuracy of current statistical machine translation systems.

The third major contribution could assist us in moving past the very superficial level of semantic representations currently in use toward much richer representations that can support the drawing of implicit inferences [2]. An example is the following sentence from a Blog document from www.warjournal.org:

While many of the weapons used by the insurgency are leftovers from the Iran-Iraq war, Iran is still providing deadly weapons such as EFPs—or Explosively Formed Projectiles.