Word Sense Disambiguation

Text, Speech and Language Technology

VOLUME 33

Eneko Agirre · Philip Edmonds
Editors

Word Sense Disambiguation
Algorithms and Applications

Springer

Eneko Agirre
University of the Basque Country
Basque Country
Spain

Philip Edmonds
Sharp Laboratories of Europe
Oxford
U.K.

ISBN 978-1-4020-4808-4 e-ISBN 978-1-4020-4809-2

Library of Congress Control Number: 2007938211

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Contents

Contributors

Eneko Agirre is an Associate Professor in the University of the Basque Country, where he is member of the IXA NLP group. He organized the Basque tasks for Senseval and coordinates the construction of the Basque WordNet and Semcor. Department of Computer Science, University of the Basque Country, Manuel de Lardizabal 1, E-20018 Donostia, Basque Country, Spain.

Paul Buitelaar is a Senior Researcher in the Language Technology Lab and co-chair of the Competence Center Semantic Web at DFKI (German Research Center for Artificial Intelligence) GmbH. He was organizer of several international workshops and has been an invited speaker at panels and workshops on topics in semantic annotation and ontology development. DFKI GmbH Language Technology Department, Stuhlsatzenhausweg 3, Saarbrücken, Germany.

Hoa Trang Dang is a Computer Scientist at the National Institute of Standards and Technology (NIST), where she coordinates evaluations of automatic question answering and summarization systems in TREC and DUC. National Institute of Standards and Technology, 100 Bureau Drive, Mailstop 8940, Gaithersburg, MD 20899-8940, U.S.A.

Philip Edmonds is a Research Scientist at Sharp Laboratories of Europe. He was chair of Senseval, 2001–2004, and is the author of the entry on lexical disambiguation in the *Elsevier Encyclopedia of Language and Linguistics, 2nd Ed.* Sharp Laboratories of Europe Limited, Oxford Science Park, Oxford OX4 4GB, United Kingdom.

Gerard Escudero is an Assistant Professor at the Universitat Politècnica de Catalunya. He was a participant in Senseval-2 and Senseval-3. He also participated in the MEANING project, funded by the EU. EUETIB, Urgell 187, E-08036 Barcelona, Catalonia, Spain.

Julio Gonzalo is an Assistant Professor at the UNED School of Computer Science. He is co-editor of the CLEF (Cross-Language Evaluation Forum) proceedings on multilingual information access published by Springer. In 2006, he is co-chair of the Programme Committee of the European

Conference on Advanced Research and Development for Digital Libraries. Dep. Lenguajes y Sistemas Informáticos, ETSI Informática – UNED, Ciudad Universitaria, c/ Juan del Rosal 16, 28040 Madrid, Spain.

Nancy Ide is a Professor of Computer Science at Vassar College and chair of the Computer Science Department. She is founder of the Text Encoding Initiative (TEI) and creator of the Corpus Encoding Standard. Currently she is directing the development of the American National Corpus. Department of Computer Science, Vassar College, 124 Raymond Avenue, Poughkeepsie, New York 12604-0520, U.S.A.

Adam Kilgarriff is Director of Lexical Computing Ltd. and Visiting Research Fellow at the University of Sussex, U.K. He works on both the theory and the practice, at the intersection of language corpora, language technologies and practical dictionary-making. Lexical Computing Ltd., 71 Freshfield Road, Brighton BN2 0BL, U.K.

Bernardo Magnini is a Senior Researcher at ITC-irst, where he coordinates the research group on Text Technologies. He is the local organizer co-chair of EACL-06, the 11th Conference of the European Chapter of the Association for Computational Linguistics. ITC-irst, Via Sommarive 18, I-38050, Povo-Trento, Italy.

Lluís Màrquez is an Associate Professor at the Polytechnical University of Catalunya (UPC). He organized two shared tasks on semantic role labeling at the Conference on Natural Language Learning (CoNLL) in 2004 and 2005, and led the team that organized the Catalan and Spanish lexical sample tasks at Senseval-3. In 2006, he will be the co-chair of the CoNLL conference. Despatx S120 - Edifici Omega, Campus Nord UPC, C/ Jordi Girona Salgado 1-3, E-08034 Barcelona, Catalonia, Spain.

David Martínez is a post-doc researcher in the NLP group of the University of Sheffield. Natural Language Processing Group, Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, United Kingdom.

Rada Mihalcea is an Assistant Professor of Computer Science at University of North Texas. She is the president of ACL SIGLEX and was a co-chair of Senseval-3. Department of Computer Science, University of North Texas, PO Box 311366, Denton, TX 76203, U.S.A.

Hwee Tou Ng is an Associate Professor of Computer Science at the National University of Singapore. He is on the editorial board of *Computational Linguistics*, was program co-chair of the ACL-2005 conference, and has served on the program committees of many past conferences including ACL, SIGIR, AAAI, and IJCAI. Department of Computer Science, School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543.

Martha Palmer is an Associate Professor in the Departments of Linguistics and Computer Science and a Faculty Fellow of the Institute of Cognitive Science at the University of Colorado at Boulder. She has been a member of the Advisory Committee for the DARPA TIDES program, chair of ACL SIGLEX and ACL SIGHAN, and is currently Past-President of the Association for Computational Linguistics. Department of Linguistics, 295 UCB - Hellems 295, Boulder, CO 80309, U.S.A.

Ted Pedersen is an Associate Professor in the Department of Computer Science at the University of Minnesota, Duluth. He is the recipient of a National Science Foundation (NSF, USA) Faculty Early Career Development (CAREER) Award. Department of Computer Science, 1114 Kirby Drive, University of Minnesota, Duluth, MN 55812, U.S.A.

Philip Resnik is an associate professor at the University of Maryland, College Park, in the Department of Linguistics and the Institute for Advanced Computer Studies. He is on the editorial board of *Cognitive Linguistics*. 1401 Marie Mount Hall, University of Maryland, College Park, MD 20742, U.S.A.

German Rigau is an Associate Professor in the Department of Computer Science of the Basque Country University. He coordinated the EU's 5th framework MEANING project. He has also participated in Senseval-2 and Senseval-3. Department of Computer Science, University of the Basque Country, Manuel de Lardizabal 1, E-20018 Donostia, Basque Country, Spain.

Mark Stevenson is a Lecturer in Computer Science at the University of Sheffield. He is author of the monograph *Word Sense Disambiguation: Combining Knowledge Sources for Sense Resolution* (2003) based on his Ph.D. thesis. Natural Language Processing Group, Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, United Kingdom.

Carlo Strapparava is a Senior Researcher at ITC-irst in the Communication and Cognitive Technologies Division. He is author of over ninety published papers on topics including artificial intelligence, natural language processing, and word sense disambiguation. ITC-irst, Via Sommarive, 18 I-38050, Povo-Trento, Italy.

M. Felisa Verdejo is Full Professor and head of the department Lenguajes y Sistemas Informáticos (LSI) at National Distance Learning University (UNED). She has been involved in several large-scale EU-funded projects such as EuroWordNet and CLEF. Dep. Lenguajes y Sistemas Informáticos, ETSI Informática – UNED, Ciudad Universitaria, c/ Juan del Rosal 16, 28040 Madrid, Spain.

Piek Vossen is CTO of Irion Technologies. He worked on several EU projects: Acquilex, Sift, EAGLES, EuroWordNet, EuroTerm, BalkaNet and MEANING and most recently on an American project to develop an Arabic wordnet. He is also founder and president of the Global Wordnet Association (GWA). Irion Technologies BV., Delftechpark 26, 2628 XH Delft, PO Box 2849, 2601 CV Delft, The Netherlands.

Yorick Wilks is Professor of Computer Science at the University of Sheffield. He is author of numerous articles and six books including *Electric Words: Dictionaries, Computers and Meanings* (1996 with Brian Slator and Louise Guthrie). He is a Fellow of the American and European Associations for Artificial Intelligence, and on the boards of some fifteen AI-related journals. Department of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom.

Foreword

Graeme Hirst

University of Toronto

Of the many kinds of ambiguity in language, the two that have received the most attention in computational linguistics are those of word senses and those of syntactic structure, and the reasons for this are clear: these ambiguities are overt, their resolution is seemingly essential for any practical application, and they seem to require a wide variety of methods and knowledge-sources with no pattern apparent in what any particular instance requires.

Right at the birth of artificial intelligence, in his 1950 paper "Computing machinery and intelligence", Alan Turing saw the ability to understand language as an essential test of intelligence, and an essential test of language understanding was an ability to disambiguate; his example involved deciding between the generic and specific readings of the phrase *a winter's day*. The first generations of AI researchers found it easy to construct examples of ambiguities whose resolution seemed to require vast knowledge and deep understanding of the world and complex inference on this knowledge; for example, *Pharmacists dispense with accuracy*. The disambiguation problem was, in a way, nothing less than the artificial intelligence problem itself. No use was seen for a disambiguation method that was less than 100% perfect; either it worked or it didn't. Lexical resources, such as they were, were considered secondary to non-linguistic common-sense knowledge of the world.

And because the methods that were developed required a resource whose eventual existence was merely hypothesized – a knowledge base containing everything a typical adult knows – and because there were no test data available, it was not possible to empirically test them or quantitatively evaluate them or their underlying ideas in any serious way. Rather, systems and methods were presented like theorems whose truth or correctness could be demonstrated by a rational argument bolstered by hand-waving and a 'toy'

demonstration: a knowledge source would be built for a few words and facts, and the system would be run on a few "interesting" constructed examples to show that it did "the right thing". This approach to evaluation was quite normal in the milieu in which this research was carried out and didn't seem to worry anyone at the time: computational linguistics had not yet achieved its empirical orientation.

Contemporary approaches have turned all that upside-down. Statistical and machine-learning methods and methodologies that have been adopted in the last decade have revolutionized our view of ambiguity resolution. It is now understood that imperfect methods that rely on rich lexical resources but limited additional knowledge have great use in the world; and that systems must undergo rigorous evaluation. The present volume demonstrates this in particular for word sense disambiguation – both the strengths and the inherent limitations of these approaches.[1] In particular, contemporary methods are less ambitious and have lower expectations. Unlike the earlier research, they don't worry about case roles, about helping a parser with attachment decisions, or about working with a semantic interpretation process aimed at a deep level of "understanding". Rather than aiming for a complete solution and hypothesizing a resource that this necessitates, they rely on an existing resource and try to see how much can be done with it. And yet they still have enormous application in NLP (see Chap. 11).

One issue that has remained constant is what kinds of information in the text may be drawn upon as cues for disambiguation, and how near in the text to the target word those cues should be. In my own early work (Hirst 1987), restrictions on communication between disambiguating processes arose from two competing principles: any particular word or structural cue for disambiguation has quite a limited sphere of influence, and yet almost anything in a text or discourse is potentially a cue for disambiguation (cf. McRoy 1992). In contemporary systems, the analogous dilemma is in the choice of features and the window size (see Chap. 8).

The other thing that hasn't changed is how hard the lexical disambiguation problem is. Many sophisticated systems struggle merely to reach the modest accuracy of simple baseline algorithms such as that of Lesk (1986) (see Chap. 5) or just choosing the most frequent sense. But what is a poor computer to do when humans themselves frequently disagree on what the correct answer is supposed to be (see Chaps. 2–4)?

Although it is an edited volume, this book is not an anthology of "recent advances" papers by individual authors on their own research, requiring

[1] A similar revolution has occurred in parsing and structural disambiguation; see Manning and Schütze (2000, Chaps. 11–12) for an overview.

each reader to synthesize a view of the overall situation in a research topic. Rather, editors Agirre and Edmonds have enlisted the leading researchers of the field to do the hard work. Each chapter of this book presents an overview and synthesis of one facet of current research. The result is a clear and well-organized presentation of the state of the art in word sense disambiguation that can be read, like a textbook, from start to finish. I commend it to you.

Graeme Hirst is the author of Semantic Interpretation and the Resolution of Ambiguity *(Cambridge University Press, 1987), which presents an integrated theory of lexical disambiguation, structural disambiguation, and semantic interpretation.*

References

Hirst, Graeme. 1987. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press.

Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of SIGDOC-86: 5th International Conference on Systems Documentation*, Toronto, Canada, 24–26.

Manning, Christopher D. & Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

McRoy, Susan. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 19(1):1–30.

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind*, 59:433–460. Reprinted in: Stuart Shieber, ed. 2004. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. Cambridge, MA: The MIT Press.

Preface

Word sense disambiguation is a core research problem in computational linguistics, which was recognized at the beginning of the scientific interest in machine translation and artificial intelligence. And yet no book has been fully devoted to review the wide variety of approaches to solving the problem. The time is right for such a book.

This book had its genesis over five years ago when Nancy Ide, series co-editor of then Kluwer's, now Springer's, *Text, Speech, and Language Technology* series, approached us with the project. Word sense disambiguation is an active and quickly progressing research field, so we thought it far more beneficial to the research community if we were to enlist the main experts to each give their own view of the field.

Being the first major book on the topic, and with the hope of it becoming the definitive reference, we endeavoured to fashion a coherent, consistent, critical, and readable survey of the current state of the art. We started by sketching an overview of the main topics that should be covered, and then approached experts in the field with desiderata for each chapter. We requested that authors give a general overview of their topic and proceed with a thorough exposition of the theory, methodology, algorithms, critical analysis, experimentation, results, and open issues. We are indebted to all of the authors, who worked with us most patiently.

The manuscript has taken time to produce, having been through numerous reviews and revisions along the way. Many difficult decisions were made in the attempt to best embrace all of the important research in the field, and to keep up with new developments. We apologize if we have missed something.

Please visit the book website, www.wsdbook.org, for the latest information updates, and a book search interface.

Word sense disambiguation is a fascinating topic; we hope you enjoy reading this book as much as we did creating it!

Acknowledgments

First all, it is the chapter authors who created the book; we thank them all for exceeding their remits and for their patience during the lengthy reviewing cycles. We owe the existence of this book to Nancy Ide, the series co-editor.

A few people gave us encouragement and feedback at various stages about the content and organization of the book. We are grateful to Robert Dale, David Farwell, Graeme Hirst, Eduard Hovy, Inderjeet Mani, Pete Whitelock, and David Yarowsky.

Three anonymous reviewers helped us get around the weak points. We also thank the team at Kluwer and Springer for their support: Tamara Welschot, Jacqueline Bergsma, Helen van der Stelt, and Jolanda Voogd (Associate Publishing Editor).

Phil was supported by Sharp Laboratories of Europe, and Eneko by the IXA research group.

Finally, this book is dedicated to our families, who had to sacrifice their time with us for the sake of this book.

Phil Edmonds and Eneko Agirre
27 January 2006