Word Sense Disambiguation
## Contents

Contributors ........................................................................................................... xiii

Foreword ................................................................................................................... xvii

Preface ................................................................................................................... xxi

1 Introduction ........................................................................................................ 1

*Eneko Agirre and Philip Edmonds*

1.1 Word Sense Disambiguation ................................................................. 1
1.2 A Brief History of WSD Research ....................................................... 4
1.3 What is a Word Sense? ....................................................................... 8
1.4 Applications of WSD ........................................................................ 10
1.5 Basic Approaches to WSD ................................................................. 12
1.6 State-of-the-Art Performance ............................................................ 14
1.7 Promising Directions ........................................................................ 15
1.8 Overview of this Book ...................................................................... 19
1.9 Further Reading ................................................................................. 21

References ............................................................................................................ 22

2 Word Senses ..................................................................................................... 29

*Adam Kilgarriff*

2.1 Introduction ......................................................................................... 29
2.2 Lexicographers ................................................................................... 30
2.3 Philosophy ........................................................................................... 32

2.3.1 Meaning is Something You Do .................................................... 32
2.3.2 The Fregean Tradition and Reification ......................................... 33
2.3.3 Two Incompatible Semantics? ..................................................... 33
2.3.4 Implications for Word Senses ....................................................... 34
2.4 Lexicalization ...................................................................................... 35
2.5 Corpus Evidence ................................................................................ 39

2.5.1 Lexicon Size ............................................................................... 41
2.5.2 Quotations .................................................................................. 42
2.6 Conclusion ............................................................................................ 43
2.7 Further Reading ................................................................................ 44

Acknowledgments ............................................................................................... 45

References ............................................................................................................ 45
3 Making Sense About Sense ................................................................. 47
Nancy Ide and Yorick Wilks
  3.1 Introduction .................................................................................... 47
  3.2 WSD and the Lexicographers ............................................................... 49
  3.3 WSD and Sense Inventories ................................................................. 51
  3.4 NLP Applications and WSD ............................................................... 55
  3.5 What Level of Sense Distinctions Do We Need
     for NLP, If Any? ...................................................................... 58
  3.6 What Now for WSD? ................................................................. 64
  3.7 Conclusion ...................................................................................... 68
References ............................................................................................ 68

4 Evaluation of WSD Systems ................................................................ 75
Martha Palmer, Hwee Tou Ng, Hoa Trang Dang
  4.1 Introduction .................................................................................... 75
    4.1.1 Terminology ............................................................................. 76
    4.1.2 Overview ............................................................................... 80
  4.2 Background ................................................................................... 81
    4.2.1 WordNet and Semcor .............................................................. 81
    4.2.2 The Line and Interest Corpora ................................................. 83
    4.2.3 The DSO Corpus ..................................................................... 84
    4.2.4 Open Mind Word Expert ......................................................... 85
  4.3 Evaluation Using Pseudo-Words ....................................................... 86
  4.4 Senseval Evaluation Exercises ....................................................... 86
    4.4.1 Senseval-1................................................................................ 87
    4.4.2 Senseval-2................................................................................ 88
    4.4.3 Evaluation and Scoring............................................................. 88
    4.4.4 English All-Words Task ............................................................ 89
    4.4.5 English Lexical Sample Task ................................................... 89
    4.4.6 Comparison of Tagging Exercises ............................................ 91
  4.5 Sources of Inter-Annotator Disagreement ...................................... 92
  4.6 Granularity of Sense: Groupings for WordNet............................... 95
    4.6.1 Criteria for WordNet Sense Grouping ....................................... 96
    4.6.2 Analysis of Sense Grouping ....................................................... 97
  4.7 Senseval-3 ....................................................................................... 98
  4.8 Discussion ....................................................................................... 99
References .......................................................................................... 102

5 Knowledge-Based Methods for WSD ................................................. 107
Rada Mihalcea
  5.1 Introduction .................................................................................. 107
  5.2 Lesk Algorithm ............................................................................. 108
    5.2.1 Variations of the Lesk Algorithm ............................................. 110
    Simulated Annealing ................................................................ 110
    Simplified Lesk Algorithm ....................................................... 111
5.3 Semantic Similarity ......................................................................114
  5.3.1 Measures of Semantic Similarity...........................................114
  5.3.2 Using Semantic Similarity Within a Local Context ..........117
  5.3.3 Using Semantic Similarity Within a Global Context.............118
5.4 Selectional Preferences................................................................119
  5.4.1 Preliminaries: Learning Word-to-Word Relations ............120
  5.4.2 Learning Selectional Preferences ........................................120
  5.4.3 Using Selectional Preferences .............................................122
5.5 Heuristics for Word Sense Disambiguation ............................123
  5.5.1 Most Frequent Sense ............................................................123
  5.5.2 One Sense Per Discourse ....................................................124
  5.5.3 One Sense Per Collocation ..........................124
5.6 Knowledge-Based Methods at Senseval-2 ............................125
5.7 Conclusions ..................................................................................126
References ..........................................................................................127

6 Unsupervised Corpus-Based Methods for WSD ......................133
Ted Pedersen
  6.1 Introduction ..................................................................................133
    6.1.1 Scope .....................................................................................134
    6.1.2 Motivation .............................................................................136
        Distributional Methods..........................................................137
        Translational Equivalence .....................................................139
    6.1.3 Approaches ............................................................................140
6.2 Type-Based Discrimination ..........................................................141
    6.2.1 Representation of Context .....................................................142
    6.2.2 Algorithms .............................................................................145
        Latent Semantic Analysis (LSA) ...............................................146
        Hyperspace Analogue to Language (HAL) ..............................147
        Clustering By Committee (CBC) ............................................148
7 Discussion........................................................................................150
  6.2.3 Discussion..............................................................................150
6.3 Token-Based Discrimination ........................................................150
    6.3.1 Representation of Context .....................................................151
    6.3.2 Algorithms .............................................................................151
        Context Group Discrimination ............................................152
        McQuitty’s Similarity Analysis ..............................................154
    6.3.3 Discussion..............................................................................157
6.4 Translational Equivalence ............................................................158
    6.4.1 Representation of Context .....................................................159
    6.4.2 Algorithms .............................................................................159
    6.4.3 Discussion..............................................................................160
6.5 Conclusions and the Way Forward ........................................... 161
Acknowledgments ............................................................................ 162
References ........................................................................................ 162

7 Supervised Corpus-Based Methods for WSD ................................. 167

Lluís Márquez, Gerard Escudero, David Martínez, German Rigau

7.1 Introduction to Supervised WSD .................................................. 167
7.1.1 Machine Learning for Classification ..................................... 168
An Example on WSD ...................................................................... 170
7.2 A Survey of Supervised WSD ..................................................... 171
7.2.1 Main Corpora Used ............................................................... 172
7.2.2 Main Sense Repositories ....................................................... 173
7.2.3 Representation of Examples by Means of Features ............... 174
7.2.4 Main Approaches to Supervised WSD .................................. 175
Probabilistic Methods ................................................................... 175
Methods Based on the Similarity of the Examples .................... 176
Methods Based on Discriminating Rules ............................. 177
Methods Based on Rule Combination .................................. 179
Linear Classifiers and Kernel-Based Approaches ................. 179
Discourse Properties: The Yarowsky Bootstrapping Algorithm ................................................................. 181
7.2.5 Supervised Systems in the Senseval Evaluations ................. 183
7.3 An Empirical Study of Supervised Algorithms for WSD .......... 184
7.3.1 Five Learning Algorithms Under Study ............................ 185
Naïve Bayes (NB) .................................................................... 185
Exemplar-Based Learning (kNN) ............................................. 186
Decision Lists (DL) ............................................................... 187
AdaBoost (AB) .................................................................... 187
Support Vector Machines (SVM) ........................................... 189
7.3.2 Empirical Evaluation on the DSO Corpus ............................ 190
Experiments ........................................................................... 191
7.4 Current Challenges of the Supervised Approach .................... 195
7.4.1 Right-Sized Training Sets ..................................................... 195
7.4.2 Porting Across Corpora ......................................................... 196
7.4.3 The Knowledge Acquisition Bottleneck ............................... 197
Automatic Acquisition of Training Examples ...................... 198
Active Learning ....................................................................... 199
Combining Training Examples from Different Words ......... 199
Parallel Corpora ..................................................................... 200
7.4.4 Bootstrapping ..................................................................... 201
7.4.5 Feature Selection and Parameter Optimization ............ 202
7.4.6 Combination of Algorithms and Knowledge Sources ....... 203
7.5 Conclusions and Future Trends ............................................... 205
8 Knowledge Sources for WSD .................................................................217
Eneko Agirre and Mark Stevenson
8.1 Introduction ..................................................................................217
8.2 Knowledge Sources Relevant to WSD .........................................218
  8.2.1 Syntactic ................................................................................219
    Part of Speech (KS 1) .................................................................219
    Morphology (KS 2) ..................................................................219
    Collocations (KS 3) ..................................................................220
    Subcategorization (KS 4) .........................................................220
  8.2.2 Semantic ................................................................................220
    Frequency of Senses (KS 5) .......................................................220
    Semantic Word Associations (KS 6) ......................................221
    Selectional Preferences (KS 7) .................................................221
    Semantic Roles (KS 8) ............................................................222
  8.2.3 Pragmatic/Topical..................................................................222
    Domain (KS 9) ........................................................................222
    Topical Word Association (KS 10) .........................................222
    Pragmatics (KS 11) ..................................................................223
8.3 Features and Lexical Resources....................................................223
  8.3.1 Target-Word Specific Features..............................................224
  8.3.2 Local Features .......................................................................225
  8.3.3 Global Features .....................................................................227
8.4 Identifying Knowledge Sources in Actual Systems .....................228
  8.4.1 Senseval-2 Systems ...............................................................229
  8.4.2 Senseval-3 Systems ...............................................................231
8.5 Comparison of Experimental Results ...........................................231
  8.5.1 Senseval Results ....................................................................232
  8.5.2 Yarowsky and Florian (2002)................................................233
  8.5.3 Lee and Ng (2002).................................................................234
  8.5.4 Martínez et al. (2002) ............................................................237
  8.5.5 Agirre and Martínez (2001a) .................................................238
  8.5.6 Stevenson and Wilks (2001)..................................................240
8.6 Discussion.....................................................................................242
8.7 Conclusions ..................................................................................245
Acknowledgments ..............................................................................246
References ..........................................................................................247

9 Automatic Acquisition of Lexical Information and Examples ......253
Julio Gonzalo and Felisa Verdejo
9.1 Introduction ..................................................................................253
9.2 Mining Topical Knowledge About Word Senses .......................254
9.2.1 Topic Signatures ................................................................. 255
9.2.2 Association of Web Directories to Word Senses ............ 257
9.3 Automatic Acquisition of Sense-Tagged Corpora ............. 258
  9.3.1 Acquisition by Direct Web Searching ......................... 258
  9.3.2 Bootstrapping from Seed Examples ......................... 261
  9.3.3 Acquisition via Web Directories ................................. 263
  9.3.4 Acquisition via Cross-Language Evidence ................. 264
  9.3.5 Web-Based Cooperative Annotation ......................... 268
9.4 Discussion ............................................................................ 269
Acknowledgments ..................................................................... 271
References ................................................................................ 272

10 Domain-Specific WSD ............................................................. 275
Paul Buitelaar, Bernardo Magnini, Carlo Strapparava, Piek Vossen
  10.1 Introduction ..................................................................... 275
  10.2 Approaches to Domain-Specific WSD ......................... 277
    10.2.1 Subject Codes .............................................................. 277
    10.2.2 Topic Signatures and Topic Variation .................... 282
      Topic Signatures ............................................................... 282
      Topic Variation ............................................................... 283
    10.2.3 Domain Tuning ............................................................. 284
      Top-down Domain Tuning ............................................... 285
      Bottom-up Domain Tuning .............................................. 285
  10.3 Domain-Specific Disambiguation in Applications .......... 288
    10.3.1 User-Modeling for Recommender Systems ............ 288
    10.3.2 Cross-Lingual Information Retrieval .................... 289
    10.3.3 The MEANING Project ............................................ 292
  10.4 Conclusions ...................................................................... 295
References ................................................................................ 296

11 WSD in NLP Applications ...................................................... 299
Philip Resnik
  11.1 Introduction ..................................................................... 299
  11.2 Why WSD? ..................................................................... 300
    Argument from Faith ......................................................... 300
    Argument by Analogy ....................................................... 301
    Argument from Specific Applications .............................. 302
  11.3 Traditional WSD in Applications ................................. 303
    11.3.1 WSD in Traditional Information Retrieval ............ 304
    11.3.2 WSD in Applications Related to Information Retrieval .. 307
      Cross-Language IR ......................................................... 308
      Question Answering ....................................................... 309
      Document Classification ............................................... 312
    11.3.3 WSD in Traditional Machine Translation ............... 313
**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

**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


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
1 Introduction

Eneko Agirre¹ and Philip Edmonds²

¹University of the Basque Country
²Sharp Laboratories of Europe Limited

1.1 Word Sense Disambiguation

Anyone who gets the joke when they hear a pun will realize that lexical ambiguity is a fundamental characteristic of language: Words can have more than one distinct meaning. So why is it that text doesn’t seem like one long string of puns? After all, lexical ambiguity is pervasive. The 121 most frequent English nouns, which account for about one in five word occurrences in real text, have on average 7.8 meanings each (in the Princeton WordNet (Miller 1990), tabulated by Ng and Lee (1996)). But the potential for ambiguous readings tends to go completely unnoticed in normal text and flowing conversation. The effect is so strong that some people will even miss a pun (a real ambiguity) obvious to others. Words may be polysemous in principle, but in actual text there is very little real ambiguity – to a person.

Lexical disambiguation in its broadest definition is nothing less than determining the meaning of every word in context, which appears to be a largely unconscious process in people. As a computational problem it is often described as “AI-complete”, that is, a problem whose solution presupposes a solution to complete natural-language understanding or common-sense reasoning (Ide and Véronis 1998).

In the field of computational linguistics, the problem is generally called word sense disambiguation (WSD), and is defined as the problem of computationally determining which “sense” of a word is activated by the use of the word in a particular context. WSD is essentially a task of classification:
word senses are the classes, the context provides the evidence, and each occurrence of a word is assigned to one or more of its possible classes based on the evidence. This is the traditional and common characterization of WSD that sees it as an explicit process of disambiguation with respect to a fixed inventory of word senses. Words are assumed to have a finite and discrete set of senses from a dictionary, a lexical knowledge base, or an ontology (in the latter, senses correspond to concepts that a word lexicalizes). Application-specific inventories can also be used. For instance, in a machine translation (MT) setting, one can treat word translations as word senses, an approach that is becoming increasingly feasible because of the availability of large multi-lingual parallel corpora that can serve as training data. The fixed inventory of traditional WSD reduces the complexity of the problem, making it tractable, but alternatives exist, as we will see below.

WSD has obvious relationships to other fields such as lexical semantics, whose main endeavour is to define, analyze, and ultimately understand the relationships between “word”, “meaning”, and “context”. But even though word meaning is at the heart of the problem, WSD has never really found a home in lexical semantics. It could be that lexical semantics has always been more concerned with representational issues (see, for example, Lyons 1995) and models of word meaning and polysemy so far too complex for WSD (Cruse 1986; Ravin and Leacock 2000). And so, the obvious procedural or computational nature of WSD paired with its early invocation in the context of machine translation (Weaver 1949) has allied it more closely with language technology and thus computational linguistics. In fact, WSD has more in common with modern lexicography, with its intuitive premise that word uses group into coherent semantic units and its empirical corpus-based approaches, than with lexical semantics (Wilks et al. 1993).

The importance of WSD has been widely acknowledged in computational linguistics; some 700 papers in the ACL Anthology mention the term “word sense disambiguation”.1 Of course, WSD is not thought of as an end in itself, but as an enabler for other tasks and applications of computational linguistics and natural language processing (NLP) such as parsing, semantic interpretation, machine translation, information retrieval, text

---

1 To compare, “anaphora resolution” occurs in 438 papers; however, such statistics should not be taken too seriously. The ACL Anthology is a digital archive of research papers in computational linguistics, covering conferences and workshops from 1979 to the present, maintained by the Association for Computational Linguistics (www.aclweb.org/anthology). Our statistics were gathered in November 2005.
mining, and (lexical) knowledge acquisition. However, in counterpoint to its theoretical importance, explicit WSD has not always demonstrated benefits in real applications.

A long-standing and central debate is whether WSD should be researched as a generic or as an integrated component. In the generic setting, the WSD component is a black box encompassing an explicit process of WSD that can be dropped into any application, much like a part-of-speech tagger or a syntactic parser. The alternative is to include WSD as a task-specific “component” of a particular application in a specific domain and integrated so completely into a system that it is difficult to separate out. Research into explicit WSD, having received the bulk of effort, has progressed steadily and successfully to a point where some people now question if the upper limit in accuracy (low as it is on fine-grained sense distinctions) has been attained (Section 1.6 gives current performance levels). And yet, explicit WSD has not yet been convincingly demonstrated to have a significant positive effect on any application. Only the integrated approach has been successful, with disambiguation often occurring implicitly by virtue of other operations, for example, in the language and translation models of statistical machine translation. The former conception is easier to define, experiment with, and evaluate, and is thus more amenable to the scientific method; the latter is more applicable and puts the need for explicit WSD into question.

Despite uncertain results on real applications, the effort on explicit WSD has produced a solid legacy of research results, methodology, and insights for computational semantics. For example, local contextual features (i.e., other words near the target word) provide better evidence in general than wider topical features (Yarowsky 2000). Indeed, the role of context in WSD is much better understood: Compared to other classification tasks in NLP (such as part-of-speech tagging), WSD requires a wide range of contextual knowledge to be modeled from fixed patterns of part-of-speech tags around a topic word to syntactic relations to topical and domain associations. Each part-of-speech and even each word relies on different types of knowledge for disambiguation. For instance, nouns benefit from a wide context and local collocations, whereas verbs benefit from syntactic features. Some words can be disambiguated by a single feature in the right position, benefiting from a “discriminative” method; others require an aggregation of many features. Homographs are generally much
easier to disambiguate than polysemous words? An evaluation methodology has been defined by Senseval (Kilgarriff and Palmer 2000) and many resources in several languages are now available. Finally, for a small sample of tested words, that have sufficient training data, the performance of WSD systems is comparable to that of humans (measured as the intertagger agreement among two or more humans), as demonstrated by the recent Senseval results (see Sect. 1.6 below).

Two “spin offs” worth mentioning include the development of explicit WSD as a benchmark application for machine learning research, because of the clear problem definition and methodology, the variety of problem spaces (each word is a separate classification task), the high-dimensional feature space, and the skewed nature of word sense distributions. And second, WSD research is helping in the development of popular lexical resources such as WordNet (Fellbaum 1998; Palmer et al. 2001, 2006) and the multilingual lexicons of the MEANING project (Vossen et al. 2006).

To introduce the topic of WSD, we begin with a brief history. Then, in Section 1.3 we discuss the central theoretical issues of “word sense” and the sense inventory. In Sections 1.4–1.6 we summarize several practical aspects including applicability to NLP tasks, the three basic approaches to WSD, and current performance achievements. Finally, Section 1.7 gathers our thoughts on emerging and future research into WSD.

1.2 A Brief History of WSD Research

In order to introduce current WSD research, reported in the book, we provide here a brief review of the history of WSD research.3

WSD was first formulated as a distinct computational task during the early days of machine translation in the late 1940s, making it one of the oldest problems in computational linguistics. Weaver (1949) introduced the problem in his now famous memorandum on machine translation:

If one examines the words in a book, one at a time through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine,

---

2 For the present purposes, a homograph is a coarse-grained sense distinction between often completely unrelated meanings of the same word string (e.g., bank as a financial institution or a river side). Polysemy involves a finer-grained sense distinction in which the senses can be related in different ways (e.g., bank as a physical building or as an institution). See Section 1.3 for further details.

3 See Ide and Véronis (1998) for a more extensive history (up to 1998, of course.)
one at a time, the meaning of words. “Fast” may mean “rapid”; or it may mean “motionless”; and there is no way of telling which.

But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning …

In addition to formulating the general methodology still applied today (see also Kaplan (1950) and Reifler (1955)), Weaver acknowledged that context is crucial, and recognized the basic statistical character of the problem in proposing that “statistical semantic studies should be undertaken, as a necessary primary step.”

The 1950s then saw much work in estimating the degree of ambiguity in texts and bilingual dictionaries, and applying simple statistical models. Zipf (1949) published his “Law of Meaning” that accounts for the skewed distribution of words by number of senses, that is, that more frequent words have more senses than less frequent words in a power-law relationship; the relationship has been confirmed for the British National Corpus (Edmonds 2005). Kaplan (1950) determined that two words of context on either side of an ambiguous word was equivalent to a whole sentence of context in resolving power.

Some early work set the stage for methods still pursued today. Masterman (1957), for instance, used the headings of the categories in Roget’s International Thesaurus (Chapman 1977) to represent the different senses of a word, and then chose the heading whose contained words were most prominent in the context. Madhu and Lytle (1965) calculated sense frequencies of words in different domains – observing early on that domain constrains sense – and then applied Bayes formula to choose the most probable sense given a context.

Early researchers well understood the significance and difficulty of WSD. In fact, this difficulty was one of the reasons why most of MT was abandoned in the 1960s due to the unfavorable ALPAC report (1966). For example, Bar-Hillel (1960) argued that “no existing or imaginable program will enable an electronic computer to determine that the word pen” is used in its ‘enclosure’ sense in the passage below, because of the need to model, in general, all world knowledge like, for example, the relative sizes of objects:

---

4 Zipf’s “Law of Meaning” is different from his well known “Zipf’s Law” about the power-law distribution of word frequencies.
Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

Ironically, the very “statistical semantics” that Weaver proposed might have applied in cases such as this: Yarowsky (2000) notes that the trigram in the pen is very strongly indicative of the enclosure sense, since one almost never refers to what is in a writing pen, except for ink.

WSD was resurrected in the 1970s within artificial intelligence (AI) research on full natural language understanding. In this spirit, Wilks (1975) developed “preference semantics”, one of the first systems to explicitly account for WSD. The system used selectional restrictions and a frame-based lexical semantics to find a consistent set of word senses for the words in a sentence. The idea of individual “word experts” evolved over this time (Rieger and Small 1979). For example, in Hirst’s (1987) system, a word was gradually disambiguated as information was passed between the various modules (including a lexicon, parser, and semantic interpreter) in a process he called “Polaroid Words”. “Proper” knowledge representation was important in the AI paradigm. Knowledge sources had to be handcrafted, so the ensuing knowledge acquisition bottleneck inevitably led to limited lexical coverage of narrow domains and would not scale.

The 1980s were a turning point for WSD. Large-scale lexical resources and corpora became available so handcrafting could be replaced with knowledge extracted automatically from the resources (Wilks et al. 1990). Lesk’s (1986) short but extremely seminal paper used the overlap of word sense definitions in the Oxford Advanced Learner’s Dictionary of Current English (OALD) to resolve word senses. Given two (or more) target words in a sentence, the pair of senses whose definitions have the greatest lexical overlap are chosen (see Chap. 5 (Sect. 5.2)). Dictionary-based WSD had begun and the relationship of WSD to lexicography became explicit. For example, Guthrie et al. (1991) used the subject codes (e.g., Economics, Engineering, etc.) in the Longman Dictionary of Contemporary English (LDOCE) (Procter 1978) on top of Lesk’s method. Yarowsky (1992) combined the information in Roget’s International Thesaurus with co-occurrence data from large corpora in order to learn disambiguation rules for Roget’s classes, which could then be applied to words in a manner reminiscent of Masterman (1957) (see Chap. 10 (Sect. 10.2.1)). Although dictionary methods are useful for some cases of word sense ambiguity (such as homographs), they are not robust since dictionaries lack complete coverage of information on sense distinctions.

The 1990s saw three major developments: WordNet became available, the statistical revolution in NLP swept through, and Senseval began.
WordNet (Miller 1990) pushed research forward because it was both computationally accessible and hierarchically organized into word senses called synsets. Today, English WordNet (together with wordnets for other languages) is the most-used general sense inventory in WSD research.

Statistical and machine learning methods have been successfully applied to the sense classification problem. Today, methods that train on manually sense-tagged corpora (i.e., supervised learning methods) have become the mainstream approach to WSD, with the best results in all tasks of the Senseval competitions. Weaver had recognized the statistical nature of the problem as early as 1949 and early corpus-based work by Weiss (1973), Kelley and Stone (1975), and Black (1988) presaged the statistical revolution by demonstrating the potential of empirical methods to extract disambiguation clues from manually-tagged corpora. Brown et al. (1991) were the first to use corpus-based WSD in statistical MT.

Before Senseval, it was extremely difficult to compare and evaluate different systems because of disparities in test words, annotators, sense inventories, and corpora. For instance, Gale et al. (1992:252) noted that “the literature on word sense disambiguation fails to offer a clear model that we might follow in order to quantify the performance of our disambiguation algorithms,” and so they introduced lower bounds (choosing the most frequent sense) and upper bounds (the performance of human annotators). However, these could not be used effectively until sufficiently large test corpora were generated. Senseval was first discussed in 1997 (Resnik and Yarowsky 1999; Kilgarriff and Palmer 2000) and now after hosting three evaluation exercises has grown into the primary forum for researchers to discuss and advance the field. Its main contribution was to establish a framework for WSD evaluation that includes standardized task descriptions and an evaluation methodology. It has also focused research, enabled scientific rigor, produced benchmarks, and generated substantial resources in many languages (e.g., sense-annotated corpora), thus enabling research in languages other than English.

Recently, at the Senseval-3 workshop (Mihalcea and Edmonds 2004) there was a general consensus (and a sense of unease) that the traditional explicit WSD task, so effective at driving research, had reached a plateau and was not likely to lead to fundamentally new research. This could indicate the need to look for new research directions in the field, some of which may already be emerging, for instance the use of parallel bilingual corpora. Section 1.7 explores the emerging research, but let’s first review the issue at the center of it all: word senses.
1.3 What is a Word Sense?

Word meaning is in principle infinitely variable and context sensitive. It does not divide up easily into distinct sub-meanings or senses. Lexicographers frequently discover in corpus data loose and overlapping word meanings, and standard or conventional meanings extended, modulated, and exploited in a bewildering variety of ways (Kilgarriff 1997; Hanks 2000; also Chap. 2). In lexical semantics, this phenomenon is often addressed in theories that model sense extension and semantic vagueness, but such theories are at a very early stage in explaining the complexities of word meaning (e.g., Cruse 1986; Tuggy 1993; Lyons 1995).

“Polysemy” means to have multiple meanings. It is an intrinsic property of words (in isolation from text), whereas “ambiguity” is a property of text. Whenever there is uncertainty as to the meaning that a speaker or writer intends, there is ambiguity. So, polysemy indicates only potential ambiguity, and context works to remove ambiguity.

At a coarse grain a word often has a small number of senses that are clearly different and probably completely unrelated to each other, usually called homographs. Such senses are just “accidentally” collected under the same word string. As one moves to finer-grained distinctions the coarse-grained senses break up into a complex structure of interrelated senses, involving phenomena such as general polysemy, regular polysemy, and metaphorical extension. Thus, most sense distinctions are not as clear as the distinction between bank as ‘financial institution’ and bank as ‘river side’. For example, bank as financial institution splits into the following cloud of related senses: the company or institution, the building itself, the counter where money is exchanged, a fund or reserve of money, a money box (piggy bank), the funds in a gambling house, the dealer in a gambling house, and a supply of something held in reserve (blood bank) (WordNet 2.1).

Even rare and seemingly innocuous words such as quoin offer a rich structure of meanings. The American Heritage Dictionary of the English Language lists three related noun-senses: the outer angle or corner of a wall, a brick forming such an angle (a cornerstone), and a wedge-shaped block. As a verb, it can mean to build a corner with distinctive blocks, or, in the printing domain, to secure metal type with a quoin.

Given the range of sense distinctions in examples such as these, which represent the norm, one might start to wonder if the very idea of word-sense is suspect. Some argue that task-independent senses simply cannot be enumerated in a list (Kilgarriff 1997; others that words are monosemous, having
Introduction

a have only a single, abstract meaning (Ruhl 1989). And perhaps the only
tenable position is that a word must have a different meaning in each dis-
tinct context in which it occurs. But a strong word-in-context position
ignores the intuition that word usages seem to cluster together into coherent
sets, which could be called senses, even if the sets cannot be satisfactorily
described or labeled. The work on sense discovery or induction gives some
empirical evidence for this intuition, however such “senses” are more aptly
called “word uses” (see Chap. 6 (Sect. 6.3)).

Concerns about the theoretical, linguistic, or psychological reality of
word senses notwithstanding, the field of WSD has successfully estab-
lished itself by largely ignoring them, much as lexicographers do in order
to produce dictionaries. Except, Kilgarriff (Chap. 2) suggests that it
is time to take notice.

In practice, the need for a sense inventory has driven WSD research. In
the common conception, a sense inventory is an exhaustive and fixed list
of the senses of every word of concern in an application. The nature of the
sense inventory depends on the application, and the nature of the disam-
biguation task depends on the inventory. The three Cs of sense inventories
are: clarity, consistency, and complete coverage of the range of meaning
distinctions that matter. Sense granularity is actually a key consideration:
too coarse and some critical senses may be missed, too fine and unneces-
sary errors may occur. For example, the ambiguity of mouse (animal or
device) is not relevant in English-Basque machine translation, where sagu
is the only translation, but is relevant in (English and Basque) information
retrieval. The opposite is true of sister, which is translated differentiy into
Basque depending on the gender of the other sibling: ahizpa for ‘sister of a
girl’ and arreba for ‘sister of a boy’. In fact, Ide and Wilks (Chap. 3) argue
that coarse-level distinctions are the only ones that humans and machines
can reliably discriminate (and that they are the distinctions of concern to
applications). There is evidence (see Chap. 4) that if senses are too fine or
unclear, human annotators also have difficulty assigning them.

The “sense inventory” has been the most contentious issue in the WSD
community, and it surfaced during the formation of Senseval, which re-
quired agreement on a common standard. The main inventories used in
English research have included LDOCE, Roget’s International Thesaurus,
Hector, and WordNet. For other languages a variety of dictionaries have
been used, together with local WordNet versions. Each resource has its
pros and cons, which will become clear throughout the book (especially
Chaps. 2, 3, and 4). For example, Hector (Atkins 1991) is lexicographi-
cally sound and detailed, but lacks coverage; LDOCE has subject codes
and a structure such that homographs are part-of-speech-homogeneous, but is not freely available; WordNet is an open and very popular resource, but is too fine-grained in many cases. Senseval eventually settled on WordNet, mainly because of its availability and coverage. Of course, this choice sidesteps the greater debate of explicit versus implicit WSD, which brings the challenge that entirely different kinds of inventory would be required for applications such as MT (translation equivalences) and IR (induced clusters of usages).

### 1.4 Applications of WSD

Machine translation is the original and most obvious application for WSD but disambiguation has been considered in almost every NLP application, and is becoming increasingly important in recent areas such as bioinformatics and the Semantic Web.

**Machine translation (MT).** WSD is required for lexical choice in MT for words that have different translations for different senses and that are potentially ambiguous within a given domain (since non-domain senses could be removed during lexicon development). For example, in an English-French financial news translator, the English noun *change* could translate to either *changement* (‘transformation’) or *monnaie* (‘pocket money’). In MT, the senses are often represented directly as words in the target language. However, most MT models do not use explicit WSD. Either the lexicon is pre-disambiguated for a given domain, hand-crafted rules are devised, or WSD is folded into a statistical translation model (Brown et al. 1991).

**Information retrieval (IR).** Ambiguity has to be resolved in some queries. For instance, given the query “depression” should the system return documents about illness, weather systems, or economics? A similar problem arises for proper nouns such as *Raleigh* (bicycle, person, city, etc.). Current IR systems do not use explicit WSD, and rely on the user typing enough context in the query to only retrieve documents relevant to the intended sense (e.g., “tropical depression”). Early experiments suggested that reliable IR would require at least 90% disambiguation accuracy for explicit WSD to be of benefit (Sanderson 1994). More recently, WSD has been shown to improve cross-lingual IR and document classification (Vossen et al. 2006; Bloehdorn and Hotho 2004; Clough and Stevenson 2004). Besides document classification and cross-lingual IR, related