Book Reviews

Word Sense Disambiguation: Algorithms and Applications

Eneko Agirre and Philip Edmonds (editors)
(University of the Basque Country and Sharp Laboratories of Europe)

Dordrecht: Springer (Text, speech, and language technology series, edited by Nancy Ide and Jean Véronis, volume 33), 2006, xxii+364 pp; hardbound, ISBN 1-4020-4804-4, $169.00, €129.95

Reviewed by
Diana McCarthy
University of Sussex

Word sense disambiguation (WSD), the tagging of words in context with labels indicating the sense in which the words are used, has become an increasingly popular area of computational linguistics research. This is particularly due to the SENSEVAL evaluation exercises which created standard data sets for the task. This book gives a thorough overview of current WSD techniques and performance of systems on these data sets, as well as a brief history of the field and some truly insightful discussions on potential developments for the future.

As Hirst points out in the Foreword, the book is a collection of summaries written by leading experts in the field rather than an anthology of authors' own work, though because these are experts there is naturally a good representation of their work. As well as surveys of existing material, there are analyses which will be of interest even to those familiar with the WSD literature. The book is an extremely useful resource for information that has not been collated elsewhere and additionally contains material that has not been published elsewhere, such as personal communication references and findings from project reports. There are some excellent discussions on the hot topics of the field. One primary focus for discussion is the choice of sense inventory for the tagging. This issue is so fundamental to the WSD community (Hanks 2000; Ide and Fellbaum 2006) that it permeates many of the chapters and relates also to another hot topic: the relevance of the task for applications. In the Introduction, the co-editors observe that although most work strives to provide useful technology from an engineering perspective, there is also scope in pursuing WSD for gains in theoretical computational semantics.

The choice of the inventory is the focus for the two chapters that follow the Introduction. Chapter 2, by Kilgarriff, looks at the notion of word sense using evidence collected from his work on the match (or mismatches) between defined senses and corpus evidence. He argues that although sense inventories might be a useful place to start, they are constructed under the pragmatic considerations of lexicographers. It should therefore not be assumed that they have cognitive validity or meet the requirements of an NLP system. Kilgarriff looks at the concept of word sense in the context of Fregean versus Gricean semantics. He argues that the Gricean framework focusing on the intended usage fits better with the spectrum of contexts of a word than trying to isolate the truth values of these contexts in a Fregean setting.

Chapter 3, by Ide and Wilks, continues the argument of appropriateness of lexicographer sense distinctions from a practical viewpoint. They argue that we would be better off abandoning the fine-grained test material that we have been focusing on and

starting from a coarser-grained level that humans and systems can more reliably discriminate. Once we are closer to 100% accuracy, we can then see if this improves a given NLP application, rather than the somewhat futile situation we have at present where system performance is too poor to be sure if WSD is beneficial to an end application. The issue remains of which word senses to use. Ide and Wilks discuss using evidence for distinctions from cross-linguistic and psycholinguistic studies or the etymology distinctions captured by lexicographers as homonyms.

Chapter 4, by Palmer, Ng, and Dang, is essential reading for those needing to get to grips with the standard evaluation data sets, particularly those created for the SENSEVAL exercises. They provide thorough descriptions of the procedures involved in producing the data sets for SENSEVAL-1 and -2 and a brief overview of SENSEVAL-3 (the book was written around the time of the latter). There are details on the range of tasks in these evaluation exercises, and the core methodology is described specifically with respect to English. There is a bias towards description of the English all-words and verbal lexical sample tasks due to the role of the first and third authors in the construction of these resources. The methods for producing the coarse-grained groups that were used for scoring the SENSEVAL-2 verbs are clearly and succinctly described. The involvement of Palmer and Dang in the construction of these resources (Palmer, Dang, and Fellbaum 2007) makes them well qualified to provide a very useful overview of this valuable work on verb classes. The methods for producing coarse senses for other parts of speech (PoS) are unfortunately not available in the SENSEVAL-2 literature.

The bulk of the book, Chapters 5–10, is devoted to WSD techniques. As the editors point out, there is inevitably some overlap. Classification of approaches and division of the material is not straightforward because topics are very much interleaved and the field abounds with hybrid systems. Some duplication could perhaps have been avoided, for example the repetition of factual material on data sets and evaluations, but the overlap is not significant and some is beneficial because it makes each chapter self-contained, reinforces important issues, and provides different perspectives on the same material.

Chapters 5–7 describe WSD methods according to three categories: knowledge-based, unsupervised, and supervised. In Chapter 5, Mihalcea gives a clear exposition of methods that use "knowledge" for disambiguation. This knowledge is typically information coded manually in a given inventory (often WordNet), though the term "knowledge-based" is also used here and elsewhere for hybrid approaches that use the predefined information to structure knowledge acquired from corpus data. Heuristics are included in this chapter, some of which rely on sense-tagged data, so they might well have been placed with the supervised methods in Chapter 7.

The division between unsupervised and supervised systems in Chapters 6 and 7 captures the difference between systems that distinguish senses according to evidence from raw data and those that make distinctions according to a predefined inventory. This is a very important distinction considering the issues with predefined inventories raised by Kilgarriff. In the WSD community, however, there is ambiguity in the term *unsupervised*, which is widely used in the WSD literature, and particularly the SENSEVAL proceedings, for systems that do not use hand-tagged data (even though they use predefined senses). The reader is warned of the ambiguity in many places but because the ambiguity is present in the book from the Introduction, it might have been worth making this ambiguity explicit from the start. In Chapter 6, Pedersen reserves the terminology for systems that discover senses automatically from data. So far such work has focused on sense discovery, with a few exceptions, such as Schütze (1998), who applied induced senses to the task of disambiguation and information retrieval. The

full potential of these unsupervised systems has yet to be realized because evaluation on standard data sets necessitates mapping from the induced classes to whichever predefined inventory was used in the creation of the data set (Agirre et al. 2006). When evaluation can be performed on a task that doesn't prescribe a particular set of labels, these methods should benefit from renewed interest.

Chapter 7 is the longest chapter in the book, which reflects the predominance of work on supervised methods that use both predefined senses and hand-tagged data. The availability of standard data sets has made systematic comparison possible; however, such comparison is difficult because so many factors are involved and typically only a few parameters are considered by any one study. In this respect, Chapters 7 and 8 contain some very useful analysis. In Chapter 7, Márquez, Escudero, Martínez, and Rigau provide an experimental comparison of some of the supervised learning algorithms on the DSO corpus. The results are dependent on the combination of algorithm, features, and evaluation data set and because the differences are small it is hard to determine definitive winners. The comparison is followed by an excellent section explaining why systems are hitting ceilings on performance and expanding on some possibilities that have been advocated for tackling the problems.

In Chapter 8, Agirre and Stevenson categorize the various knowledge sources that feed into WSD systems, identify these sources in existing WSD systems, and collate findings on the benefits of the various categories from a number of comparative evaluations in the literature. They provide a useful set of observations on the contribution of the knowledge sources described in the chapter. Again it is apparent that although there are general trends to be found—for example that verbs may do better with specific knowledge sources, such as subcategorization, and discriminative approaches—there is no "one size fits all" even across a given PoS and the interaction between features and algorithms makes the exploration difficult. There is plenty of motivation for combining knowledge sources to get optimum results, as no one knowledge source or algorithm is a panacea; however, work is clearly needed to isolate components and to determine what works well when. The discussion in Chapter 8 is a useful step in that direction.

In Chapter 9, Gonzalo and Verdejo examine how knowledge sources needed for WSD can be acquired automatically. They look at automatic acquisition of topical knowledge about word senses and also pick up on the thread from the end of Chapter 7 on trying to provide supervised systems with sufficient sense-tagged examples using cross-lingual resources and information gleaned from the Web. They highlight research (Agirre and Martínez 2004) demonstrating the importance of determining the right sampling bias when using Web data. Another approach has been to exploit the Web community for voluntary labor in annotation tasks. Web directories also show potential for finding valuable domain-specific information, which is championed in Chapter 10 as a crucial input for WSD.

Chapter 10 is a practical chapter on the importance of domains, subjects, and topics. Buitelaar, Magnini, Strapparava, and Vossen ask whether domains are necessary for WSD and whether they are the whole story. The evidence suggests that the answer is somewhere between these two viewpoints. There is no doubt that for many words, domain information is important, though the high percentage of "factotum" (i.e., domain-independent) words implies it cannot be the whole answer. The importance of the domain issue depends on the purpose of WSD. If one were using WSD output for semantic analysis of generic text, then domain issues would not be as significant compared to a cross-lingual task operating on domain-specific text. The chapter ends

with a series of results demonstrating the potential for domain-specific WSD in various information retrieval and cross-lingual information retrieval tasks.

The book finishes with a wonderful survey by Resnik on the role of WSD in applications. If one is making claims about the potential utility of an explicit WSD module, which most WSD research does, one needs to be aware of the lack of proof for this assumption. Resnik presents the evidence in a very readable summary along with reasons why the benefits of WSD have yet to be proved. He does, however, argue that we are right to endeavor to validate our techniques by demonstrating that they have practical utility, and he speculates on some emerging applications where the WSD technologies might find a niche.

This book is an excellent overview of a buoyant research area. We are now as a community looking to broaden our horizons as we look forward to SEMEVAL, the successor of the SENSEVAL exercises. This collection serves as a thorough record of where we are now and provides some nice pointers for where we need to go. It is a great resource containing valuable reference material, helpful summaries of findings, further-reading sections, and a useful appendix on resources. There is also an index to many of the authors and algorithms cited in the book, though not all cited systems actually appear in the index. Even though the book is tailored for those new to the field, veteran WSD researchers will find the collection makes good reading with plenty of material and discussions that do not appear elsewhere. I will certainly be dipping into the book for many years to come.

References

Agirre, Eneko and David Martínez. 2004. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 25–32, Barcelona, Spain.

Agirre, Eneko, David Martínez, Oier López de Lacalle, and Aitor Soroa. 2006. Two graph-based algorithms for state-of-the-art WSD. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 585–593, Sydney, Australia.

Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities.* Senseval Special Issue, 34(1–2):205–215.

Ide, Nancy and Christiane Fellbaum, editors. 2006. *Proceedings of the EACL 06 Workshop: Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, Trento, Italy.

Palmer, Martha, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Diana McCarthy is a UK Royal Society Dorothy Hodgkin Fellow at the University of Sussex. Her research has focused on lexical semantics and she has been involved in the SENSEVAL WSD evaluation exercises since their inception as a participant and program committee member and more recently as a task organizer for SEMEVAL. She is currently on the executive board of SIGLEX, the ACL's special interest group on the lexicon. McCarthy's address is Department of Informatics, University of Sussex, Falmer, Brighton, East Sussex, BN1 9QH, UK; e-mail: dianam@sussex.ac.uk.