Mark Stevenson

Word Sense Disambiguation
The Case for Combinations of Knowledge Sources

CSLI Publications, 2003
Stanford, California
ISBN 1-57586-390-1 (pbk.), 1-57586-389-8 (hard)
xvi+175 pages

Reviewed by
Zdeněk Žabokrtský

In general, the term 'Word Sense Disambiguation' (WSD for short) is used for the process of deciding, which sense (of all the senses of a given word) is used in a given context. In NLP, the automation of this process is notoriously difficult. The book under review surveys the progress in the field and reports on an extensive novel research. The author explored the topic of WSD while working in the Natural Language Processing group of Sheffield University. The book is an extension of his Ph.D. thesis.

The book starts with foreword by Yorick Wilks (the author's adviser) and preface. The rest of the text is divided into nine chapters:

Chapter 1 – Introduction. The first chapter explains the goal of WSD and documents the motivation for it on several NLP applications. Then the classification of WSD tasks is presented: Semantic Disambiguation (the most general: without restrictions of the type of added semantic information of portion of annotated text), Semantic Tagging (all words are annotated), Sense Disambiguation (only senses from a lexicon can be added), and Sense Tagging (all words are annotated with their senses from a lexicon).

Chapter 2 – Background. State of the art in WSD is briefly outlined. It starts with Bar-Hillel's skepticism in the sixties, then several Data Based Approaches are mentioned, and finally a taxonomy of WSD algorithms is suggested.

Chapter 3 – Meaning and the Lexicon. The difference between the two traditional levels of meaning distinctions, namely between homography and polysemy, is explained. Then the evolution of dictionaries from Samuel Johnsons's A Dictionary of the English Language to the modern Machine Readable Dictionaries is briefly sketched. Three English NLP lexicons are discussed in more detail: Longman Dictionary of Contemporary English (LDOCE), Roget's Thesaurus and WordNet. The structure and the coverage of the three lexicons are compared. The modern lexicographic practice is described on the example of the COBUILD dictionary and the possibilities of the automation of the lexicographic process are discussed. Then three criticisms of machine readable dictionaries are explained (Kilgarriff's attack against the 'Bank Model', Pustejovsky's criticism of limitations of Sense Enumerated Lexicons, Kay's suggestions about an ideal abstract lexicon), but the author convincingly defends the dictionary model. In my opinion, this is the most interesting part of the first (survey) half of the book.

Chapter 4 – A Framework for Disambiguation. In this chapter, the notion of Weak Knowledge Sources is explained (sources of information which solve some instances of a problem in question but are not sufficient for solving them all), and the question how they can be systematically combined is discussed. Then two conditions for the WSD framework are assumed: 1) the algorithm has access to at least one knowledge source, 2) each word has a finite set of distinct senses. The disambiguation modules are divided into three classes: 1) Filters (filters are used when some senses can be removed from considerations with high reliability), 2) Partial Taggers (these mark the potential candidate(s)), 3) Feature Extractors (do not disambiguate, only extract specific data from the context of an ambiguous word). An application of this scheme is illustrated on two case studies: LDOCE and WordNet.

Chapter 5 – Part of Speech and Sense Tagging. The similarity of the two tasks is discussed. The contribution of the information about part of speech to the WSD task (mainly on the homograph level) is evaluated in two experiments, and is found to be significant.

Chapter 6 – Implementation. A sense tagger implemented within the framework described in Chapter 4 is presented. The complex architecture of the system is divided into the preprocessing and disambiguations parts. Preprocessing consists of tokenisation, part of speech tagging, named entity identification, simple grammatical link identification and lexical lookup. Then the following disambiguation modules are used: part of speech filter, simulated annealing, broad context and selectional restriction partial taggers and collocation extractor.

Chapter 7 – Sense Tagged Corpora. The author justifies the need for some 'gold standard' data for the evaluation purposes. He reviews several available corpora that contain annotation relevant for WSD. He discusses the cost of a virtually ideal corpus and outlines alternative approaches for acquiring more sense-annotated data, mainly via mapping between lexical resources.

Chapter 8 – Evaluation. In this chapter the evaluation methodology and the experimental setting are described, and the performance of individual partial taggers is evaluated. The obtained results (90% of ambiguous words were disambiguated on the sense level, 94% of polyhomographic words were disambiguated on the homography level) are compared to other studies.

Chapter 9 – Conclusion. Finally, the contribution of the work is summarized and some ideas for future research are outlined.

The book is well-arranged and reads well, too. Most chapters contain their own introductions and conclusions, which make the reading more comfortable. The text describing related work of other researchers can be always clearly differentiated from the contribution of the author. The survey part of the book does not completely cover the state of the art, but the reader is informed where a more comprehensive overview of WSD can be found. Implementation of a new WSD system, which is capable of using information from different knowledge sources and combines new and existing approaches to WSD, is reported. It is well documented and evaluated in detail, and interpretation of the results is suggested. The resulting performance of the reported system is high.

If I am to find also some aspect in which the book deserves criticism, then I should mention its anglocentrism. It is hard to believe that nothing worth mentioning was done in WSD and/or lexicography of other languages (for example, there are some Machine Translation systems around, where WSD for languages other than English must have already been faced 'in vivo'). And even if it is true that there is nothing notable as for other languages what was not done for English too, one would at least expect the author to comment the potential applicability of his approach outside English. As an illustration, I just randomly pick two of many undiscussed issues: a) Slavic languages (due to their inflectional nature) probably possess much lower degree of homography compared to English, which might significantly reduce the information contribution of part-of-speech tagger to the WSD task, b) it is not clear how to treat senses of German composites – 'new words' that are frequently and productively created via concatenation and are not listed in lexicons.

If I am to formulate my personal 'take-home-message' from the book, I would use the following words: The questions related to WSD were recognized within NLP in its earliest days and have been discussed since, but unfortunately no broadly accepted view on the term 'word sense' has been reached so far. However, the absence of a solid theoretical background does not prevent contemporary NLP practitioners from experimenting with (mostly off-the-shelf) lexicographical resources and automating the process of WSD with 90% agreement with a human (which I find really surprising in the light of the above fact).

In my opinion, the book will not only be an interesting reading for specialists in WSD and professional lexicographers, but – since it is written in a very clear language and contains explanations of many basic notions common to most NLP sub-domains – it could be also used as a supplementary study material for various NLP courses.