D-Lib Magazine
October 1998

ISSN 1082-9873

WordNet

blue line

"Natural language processing is essential for dealing efficiently with the large quantities of text now available online: fact extraction and summarization, automated indexing and text categorization, and machine translation. Another essential function is helping the user with query formulation through synonym relationships between words and hierarchical and other relationships between concepts. WordNet supports both of these functions and thus deserves careful study by the digital library community."

By Dagobert Soergel
ds52@umail.umd.edu

Dagobert Soergel is Professor in the College of Library and Information Services, University of Maryland and specializes in the development of classification schemes and thesauri.

WordNet
An electronic lexical database
Christiane Fellbaum, ed., with a preface by George Miller
423 pages. Illustrations, Index. Cambridge, Massachusetts and London, England:
The MIT Press, Massachusetts Institute of Technology 1998, $50 (cloth).
www-mitpress.mit.edu

WordNet. An electronic lexical database. Edited by Christiane Fellbaum, with a preface by George Miller. Cambridge, MA: MIT Press; 1998. 422 p. $50.00

This is a landmark book. For anyone interested in language, in dictionaries and thesauri, or natural language processing, the introduction, Chapters 1- 4, and Chapter 16 are must reading. (Select other chapters according to your special interests; see the chapter-by-chapter review). These chapters provide a thorough introduction to the preeminent electronic lexical database of today in terms of accessibility and usage in a wide range of applications. But what does that have to do with digital libraries? Natural language processing is essential for dealing efficiently with the large quantities of text now available online: fact extraction and summarization, automated indexing and text categorization, and machine translation. Another essential function is helping the user with query formulation through synonym relationships between words and hierarchical and other relationships between concepts. WordNet supports both of these functions and thus deserves careful study by the digital library community.

The introduction and part I, which take almost a third of the book, give a very clear and very readable overview of the content, structure, and implementation of WordNet, of what is in WordNet and what is not; these chapters are meant to replace *Five papers on WordNet* (ftp://ftp.cogsci.princeton.edu/pub/WordNet/5papers.ps), which by now are partially outdated. However, I did not throw out my copy of the *Five papers*; they give more detail and interesting discussions not found in the book. Chapter 16 provides a very useful complement; it includes a very good overview of WordNet relations (with examples and statistics) and describes possible extensions of content and structure.

Part II, about 15% of the book, describes "extensions, enhancements, and new perspectives on WordNet", with chapters on the automatic discovery of lexical and semantic relations through analysis of text, on the inclusion of information on the syntactic patterns in which verbs occur, and on formal mathematical analysis of the WordNet structure.

Part III, about half the book, deals with representative applications of WordNet, from creating a "semantic concordance" (a text corpus in which words are tagged with their proper sense), to automated word sense disambiguation, to information retrieval, to conceptual modeling. These are good examples of pure knowledge-based approaches and of approaches where statistical processing is informed by knowledge from WordNet.

As one might expect, the papers in this collection (which are reviewed individually below), are of varying quality; Chapters 5, 12, and 16 stand out. Many of the authors pay insufficient heed to the simple principle that the reader needs to understand the overall purpose of work being discussed in order to best assimilate the detail. Repeatedly, one finds small differences in performance discussed as if they meant something, when it is clear that they are not statistically significant (or at least it is not shown that they are), a problem that afflicts much of information retrieval and related research. The application papers demonstrate the many uses of WordNet but also make a number of suggestions for expanding the types of information included in WordNet to make it even more useful. They also uncover weaknesses in the content of WordNet by tracing performance problems to their ultimate causes.

The papers are tied together into a coherent whole, and that unity of the book is further enhanced by the presence of an index, which is often missing from such collections. One would wish the index a bit more detailed.. The reader wishing for a comprehensive bibliography on WordNet can find it at www.cis.upenn.edu/~josephr/wn-biblio.html (and more information on WordNet can be found at www.cogsci.princeton.edu/~wn).

The book deals with WordNet 1.5, while WordNet 1.6 was released at just about the same time. For the papers on applications it could not be otherwise, but the chapters describing WordNet might have been updated to include the 1.6 enhancements. One would have wished for a chapter or at least a prominent mention of EuroWordNet (www.let.uva.nl/~ewn), a multilingual lexical database using WordNet's concept hierarchy.

It would have been useful to enforce common use at least of key terms throughout the papers; for example, the introduction says "WordNet makes the commonly accepted distinction between *conceptual-semantic relations*, which link concepts, and *lexical relations*, which link individual words [emphasis added], yet in Chapter 1 George Miller states that "the basic *semantic* relation in WordNet is *synonymy*" [emphasis added] and in Chapter 5 Marti Hearst uses the term *lexical relationships* to refer to both *semantic relationships* and *lexical relationships*. Priss (p. 186) confuses the issue further in the following statements: "Semantic relations, such as meronymy, *synonymy*, and hyponymy, are according to WordNet *terminology* relations that are defined among synsets. They are distinguished from lexical relations, such as antonymy, which are defined among words and not among synsets." [Emphasis added]. Synonymy is, of course, a lexical relation, and terminology relations are relations between words or between words and concepts, but not relations between concepts.

This terminological confusion is indicative of a deeper problem. The authors of WordNet vacillate somewhat between a position that stays entirely in the plane of words and deals with conceptual relationships in terms of the relationships between words, and the position, commonly adopted in information science, of separating the conceptual plane from the terminological plane.

What follows is a chapter-by-chapter review. By necessity, this often gets into a discussion how things are done in WordNet, rather than just staying with the quality of the description per se.

Part I. *The lexical database* (p. 1 - 127)

George Miller's preface is well worth reading for its account of the history and background of WordNet. It also illustrates the lack of communication between fields concerned with language: Thesaurus makers could learn much from WordNet, and WordNet could learn much from thesaurus-makers. The introduction gives a concise overview of WordNet and a preview of the book. It might still make clear more explicitly the structure of WordNet: The smallest unit is the word/sense pair identified by a sense key (p. 107); word/sense pairs are linked through WordNet's basic relation, synonymy, which is expressed by grouping word/sense pairs into synonym sets or synsets; each synset represents a concept, which is often explained through a brief gloss (definition). Put differently, WordNet includes the relationship word W *designates* concept C, which is coded implicitly by including W in the synset for C. (This is made explicit

in Table 16.1.) Two words are synonyms if they have a *designates* relationship to the same concept. Synsets (concepts) are the basic building blocks for hierarchies and other conceptual structures in WordNet.

The following three chapters each deal with a type of word. In Chapter 1, *Nouns in WordNet*, George Miller presents a cogent discussion of the relations between nouns: synonymy, a lexical relation, and hyponymy (has narrower concept) and meronymy (has part), semantic relations. Hyponymy gives rise to the WordNet hierarchy. Unfortunately, it is not easy to get well-designed display of that hierarchy. Meronymy (has part) / holonymy (is part of) always are subject to confusion, not completely avoided in this chapter, which stems from ignoring the fact that *airplane has-part wing* really means "an individual object 1 which belongs to the class airplane has-part individual object 2 which belongs to the class wing". While it is true that *bird has-part wing*, it is by no means the same wing or even the same type of wing. Thus the strength of a relationship established indirectly between two concepts due to a has-part relationship to the same concept may vary widely. Miller sees meronymy and hyponymy as structurally similar, both going down. But it is equally reasonable to consider a *wing* as something more general than an *airplane* and thus construe the has-part relationship s going up, as is done in Chapter 13.

Chapter 2, *Modifiers in WordNet* by Katherine J. Miller describes how adjectives and adverbs are handled in WordNet. It is claimed that nothing like hyponymy/hypernymy is available for adjectives. However, especially for information retrieval purposes, the following definition makes sense: Adj A has narrower term (NT) Adj B if A applies whenever B applies and A and B are not synonyms (the scope of A includes the scope of B and more). For example, *red NT ruby* (and conversely, *ruby BT red*) or *large NT huge*. To the extent that WordNet includes specific color values, their hierarchy is expressed only for the noun form of the color names. Hierarchical relationships between adjectives are expressed as *similarity*. The treatment of antonymy is somewhat tortured to take account of the fact that *large vs small* and *big vs. little* are antonym pairs, but **large vs. little* or **prompt vs. slow* are not. In general, when speakers express semantic opposition, they do not use just any term for each of the opposite concepts, but they apply lexical selection rules. The simplest way to deal with this problem is to establish an explicit *semantic* relationship of opposition between *concepts* and separately record antonymy as a *lexical* relationship between *words* (in their appropriate senses). Participial adjectives are maintained in a separate file and have a pointer to the verb unless they fit easily into the cluster structure of descriptive adjectives "rather than" having a pointer to the verb. Why a participial adjective in the cluster structure cannot also have a pointer to the verb is not explained.

In Chapter 3, *A semantic network of English verbs*, Christiane Fellbaum presents a carefully reasoned analysis of the semantic relations between verbs (as presented in WordNet and as presented in alternate approaches) and of the other information for verbs, especially sentence structures, given in WordNet. Much of this discussion is based on results from psycholinguistics. The relationships between verbs are all forms of a broadly defined entailment relationship: troponymy establishes a verb hierarchy (verb A has the troponym verb B if B is a particular way of doing A, corresponding to hyponymy between nouns; the reverse relation is verb B has hypernym verb A); *entailment* in a more specific sense (to snore entails to sleep); *cause*; and *backward presupposition* (one must know before one can forget, not given in

WordNet). More refined relationship types could be defined but are not for WordNet, since they do not seem to be psychologically salient and would complicate the interface. The last is not a good reason, since detailed relationship types could be coded internally and mapped to a more general set to keep the interface simple; the additional information would be useful for specific applications. In Section 3.3.1.3 on basic level categories in verbs, the level numbers are garbled (talk is both level L+1 and level L), making the argument unclear.

In Chapter 4, *Design and implementation of the WordNet lexical database and searching software,* Randee I. Tengi lays out the technical issues. While it is hard to argue with a successful system, one might observe that this is not the most user-friendly system this reviewer has seen. The lexicographers have to memorize a set of special characters that encode relationship types (for example, ~ for hyponym, @ for hypernym); {apple, edible_fruit, @} means: apple has the hypernym edible_fruit. The interface is on a par with most online thesaurus interfaces, that is to say, poor. In particular, one cannot see a simple outline of the concept hierarchy; for many thesauri that is available in printed form. P. 108 says that "many relations being reflexive", when it should say reciprocal (such as hyponym / hypernym); a reflexive relation has a precise definition in mathematics (for all a, a R a holds).

Chapter 16, *Knowledge processing on an extended WordNet* by Sanda M. Harabagiu and Dan I. Moldovan is discussed here because it should be read here. Table 16.1 gives the best overview of the relations in WordNet, with a formal specification of the entities linked, examples, properties (symmetry, transitivity, reverse of), and statistics. Table 16.2 gives further types of relationships that can be derived from the glosses (definitions) provided in WordNet. Still further relationships can be derived by inferences on relationship chains. The application of this extended WordNet to knowledge processing (for example, marker propagation, testing a text for coherence) is less well explained.

Part II. Extensions, enhancements, and new perspectives on WordNet p. 129 - 196

Chapter 5, *Automated discovery of WordNet relations,* by Marti A. Hearst is a well-written account of using Lexico-Syntactic Pattern Extraction (LSPE) to mine text for hyponym relationships. This is important to reduce the enormous workload in the strictly intellectual compilation of the large lexical databases needed for NLP and retrieval tasks. The following examples illustrate two sample patterns (the implied hyponym relationships should be obvious):

> red algae, *such as* Gelidium
>
> bruises, broken bones, *or other* injuries

Patterns can be hand-coded and/or identified automatically. The results presented are promising. In the discussion of Automatic acquisition from corpora (p. 146, under Related work), only fairly recent work from computational linguistics is cited. Resnik, in Chapter 10, cites work back to 1957. There is also early work done in the context of information retrieval, for example, Giuliano 1965; for more references on early work see Soergel 1974, Chapter H.

Chapter 6, *Representing verb alternations in WordNet,* by Karen T. Kohl, Douglas A. Jones, Robert C. Berwick, and Naoyuki Nomura discusses a format for representing syntactic patterns of verbs in WordNet. It is clearly written for linguists and a heavy read for others. It is best to look at the appendix first to get an idea of the ultimate purpose.

Chapter 7, *The formalization of WordNet by methods of relational concept analysis,* by Uta E. Priss is somewhat of an overkill in formalization. Formal concept analysis might be a useful method, but the starkly abbreviated presentation given here is not enough to really understand it. The paper does reveal a proper analysis of meronymy, once one penetrates the formalism and the notation. The WordNet problems identified on p. 191-195 are quite obvious to the experienced thesaurus builder once the hierarchy is presented in a clear format and could in any event be detected automatically based on simple componential analysis.

Part III Applications of WordNet. p. 197-406

Chapters 8 *Building semantic concordances,* by Shari Landes, Claudia Leacock, and Randee I. Tengi and Chapter 9, *Performance and confidence in a semantic annotation task,* by Christiane Fellbaum, Joachim Grabowski, and Shari Landes both deal with manually tagging each word in a corpus (in the example, 103 passages from the Brown corpus and the complete text of Crane's *The red badge of courage*) with the appropriate WordNet sense. This is useful for many purposes: detecting new senses, obtaining statistics on sense occurrence, testing the effectiveness of correct disambiguation for retrieval or clustering of documents, and training disambiguation algorithms. The tagged corpus is available with WordNet.

Chapter 10, *WordNet and class-based probabilities,* by Philip Resnik, presents an interesting and discerning discourse on "how does one work with corpus-based statistical methods in the context of a taxonomy?" or, put differently, how does one exploit the knowledge inherent in a taxonomy to make statistical approaches to language analysis and processing (including retrieval) more effective. However, the real usefulness of the statistical approach does not appear clearly until Section 10.3, where the approach is applied to the study of selectional preferences of verbs (another example of mining semantic and lexical information from text), and thus the reader lacks the motivation to understand the sophisticated statistical modeling. The reader deeply interested in these matters might prefer to turn to the fuller version in Resnik's thesis (access from www.umiacs.umd.edu/~Resnik),

Chapters 11, *Combining local context and WordNet similarity for word sense identification,* by Claudia Leacok and Martin Chodorov and 13, *Lexical chains as representations of context for the detection and correction of malapropisms,* by Graeme Hirst and David St-Onge present methods for sense disambiguation. Chapter 11 uses a statistical approach, computing from a sense-tagged corpus (the training corpus) three distributions that measure the associations of a word in a given sense with part-of-speech tags, open-class words, and closed-class words. These distributions are then used to predict the sense of a word in arbitrary text. The key idea of the paper is very close to the basic idea of Chapter 10: exploit knowledge to improve statistical methods, in this case use knowledge on word sense similarity gleaned from WordNet to extend the usefulness of the training corpus by using the distributions for a sense A1 of word A to

disambiguate word B, one of whose senses is similar to A1. Results of experiments are inconclusive (they lack statistical significance), but he idea seems definitely worth pursuing. Hirst and St-Onge take an entirely different approach. Using semantic relations from WordNet they build what they call lexical chains (semantic chain would be a better term) of words that occur in the text. Initially there are many competing chains, corresponding to different senses of the words occurring in the text, but eventually some chains grow long and others do not, and the word senses in the long chains are selected. (Instead of chains one might use semantic networks.) Various patterns of direct and indirect relationships are defined as allowable chain links, and each pattern has a given strength. In the paper, this method is used to detect and correct malapropisms, defined here as misspellings that are valid words (and therefore not detected by a spell checker), for example, "Much of the data is available *toady* electronically" or "Lexical relations *very* in number within the text". (Incidentally, this sense of malapropism is not found in standard dictionaries, including WordNet; in a book on WordNet, one should use language carefully.) The problem of detecting misspellings that are themselves words is the same as the homonym problem: If one accepts, for example, both *today* and *toady* as variations of the word *today* and of the word *toady*, then [*today. toady*] becomes a homonym having all the senses of the two words that are misspellings of each other; that is to say, the occurrence of either word in the text may indicate any of the senses of both words. Finding the sense that fits the context selects the form correctly associated with that sense. This is an ingenious application of sense disambiguation methods to spell checking.

Chapter 14, *Temporal indexing through lexical chaining*, by Reem Al-Halimi and Rick Kazman uses a very similar approach to indexing text, an idea that seems quite good (as opposed to the writing). The texts in question are transcripts from video conferences, but that is tangential and their temporal nature is not considered in the paper. Instead of chains they build lexical trees (again, semantic tree would be a better name, and why not use more general semantic networks). A text is represented by one or more trees (presumably each tree representing a logical unit of the text). Retrieval then is based on these trees as target objects. A key point in this method, not stressed in the paper, is the word sense disambiguation that comes about in building the semantic trees. The method appears promising, but no retrieval test results are reported.

In Chapter 12, *Using WordNet for text retrieval*, Ellen M. Voorhees reports on retrieval experiments using WordNet for query term expansion and word sense disambiguation. No conclusions as to the usefulness of these methods should be drawn from the results. The algorithms for query term expansion and for word sense disambiguation performed poorly — partially due to problems in the algorithms themselves, partially due to problems in WordNet — so it is not surprising that retrieval results were poor. Later work on improved algorithms, including work by Voorhees, is cited.

In Chapter 15, *COLOR-X: Using knowledge from WordNet for conceptual modeling* [in software engineering], J. F. M. Burg and R. P. van de Riet describe the use of WordNet to clarify word senses in software requirement documents and for detecting semantic relationships between entity types and relationship types in an E-R model. The case they make is not very convincing. They also bring the art of making acronym soup to new heights, misleading the reader in the process (the chapter has nothing to do with colors).

All in all, this is a useful collection of papers and a rich source of ideas.

References

Giuliano, Vincent E. 1965 The interpretation of word associations. In Stevens, Mary E.; Giuliano, Vincent E.; Heilprin, Lawrence B., eds. Statistical association methods for mechanical documentation. Washington, DC: Government Printing Office; 1965 December. 261 p. (National Bureau of Standards Miscellaneous Publication 269)

Soergel, Dagobert. 1974. Indexing languages and thesauri: Construction and maintenance. Los Angeles, CA: Melville; 1974. 632 p., 72 fig., ca 850 ref. (Wiley Information Science Series)