Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon

Uri Zernik (editor)
(Research and Development Center, General Electric Company)

Hillsdale, NJ: Lawrence Erlbaum
Associates, 1991, ix + 429 pp.
Hardbound, ISBN 0-8058-0829-9, $69.95;
Paperbound, ISBN 0-8058-1127-3, $34.50

Reviewed by
Victor Sadler
BSO/Artificial Intelligence, Utrecht

It's taken a good three years to squeeze this collection of workshop papers through the publishing pipeline. No great matter: the state of the art in lexical acquisition is still much as it was. Editor Zernik has used the time to impose some coherence on the multi-disciplinary material by adding a substantial introduction, and at least some of the authors have cross-referenced their papers to other contributions. The effort has succeeded up to a point. But, as with most such proceedings, it is still up to the reader to fit the bits and pieces into a common frame of reference.

The papers are as follows:

1. Uri Zernik, "Introduction"
2. Paul Jacobs, "Making sense of lexical acquisition"
3. Robert Krovetz, "Lexical acquisition and information retrieval"
4. Brian Slator, "Using context for sense preference"
5. Uri Zernik, "Tagging word senses in corpus"
6. Kenneth Church, William Gale, Patrick Hanks, and Don Hindle, "Using statistics in lexical analysis"
7. Frank Smadja, "Macrocoding the lexicon with co-occurrence knowledge"
8. Nicoletta Calzolari, "Lexical databases and textual corpora: Perspectives of integration for a lexical knowledge-base"
9. C. Felbaum, D. Gross, and G. Miller, "WordNet: A lexical database organized on psycholinguistic principles"
10. Beryl Atkins and Beth Levin, "Admitting impediments"
11. Bonnie Dorr, "Conceptual basis of the lexicon in machine translation"
12. Michael Dyer, "Lexical acquisition through symbol recirculation in distributed connectionist networks"
13. P. Velardi, "Acquiring a semantic lexicon for natural language processing"
14. Lisa Braden-Harder and Włodek Zadrozny, "Lexicons for broad coverage semantics"
15. James Martin, "Representing and acquiring metaphor-based polysemy"

(In what follows, figures in parentheses refer to the above list.)

How can we use on-line sources to provide NLP systems with lexical data in at least a semi-automatic fashion? The question is a pressing one, because "the Lexicon has emerged as the major natural language processing bottleneck" (1). Most existing systems stall at unknown words. Even when a word is known, there is no guarantee that the current sense of that word is known, or that its context does not constitute an unknown idiom or compound, or that other data in the lexicon are adequate. Generation in particular requires extensive knowledge of "idiosyncratic" (7) collocations (e.g., *strong tea* or *powerful car*), which do not fit into the conventional linear conception of a lexicon. Moreover, "understanding NL requires vast amounts of background knowledge" (14). How that knowledge should be represented and keyed to lexicon entries is still an open question.

There is, of course, broad consensus about some of the basic stuff that should go into the lexicon, starting with word class. Atkins and Levin (10) offer a shortlist for verbs: for each sense, the lexicon should contain semantic class, aktionsart, and arguments; selectional restrictions on arguments; subcategorization; morphologically related words; related extended uses; related idiomatic uses; collocates; domain labels; pragmatic force; corpus citations exemplifying each feature; plus, of course, phonology and morphology.

Other papers extend or refine this list. Relative frequencies of word senses and collocates are needed (3, 14), for example, if lexical choices in natural language output are to appear "natural" (6). But of course, there is no absolute criterion for splitting an entry into "word senses" in the first place (5, 10). And collocations need to be stored as annotated grammatical structures (1) if productive use is to be made of them (*5th grade* → *6th grade(r)*, etc). Synonyms and antonyms should also be accessible (14), as well as IS-A links and other types of semantic cross-reference as represented in Princeton's WordNet system (9). Calzolari (8) sees in the addition of such links the basic difference between a lexical database and a lexical knowledge base. For representing meaning, Dorr (11) requires Schankesque conceptual structures, while Dyer (12) proposes context-based statistical associations. These two approaches reflect Velardi's (13) distinction (following Leech) between conceptual meaning and collocative meaning. Both may be needed.

Now to the crux: How can we automate the acquisition of all these data? Most contributors consider one of two possible on-line sources: machine-readable dictionaries (MRDs) or text corpora. While a few types of data can be usefully extracted from MRDs (8, 9), e.g., word senses for the purposes of information retrieval or topic identification (3, 4), most cannot be. And merging data from different MRDs is highly problematic (10). Collocations (1, 6, 7), subcategorization (8), and any semantic knowledge beyond basic semantic features (14) are more readily obtainable from corpora (13), and some useful tools are described (5, 6, 7). In either case, the consensus is that human intervention or post-editing is inevitable (2, 6, 13). On-line texts also require linguistic pre-processing (1, 6) before useful data can be extracted, but given the limitations of state-of-the-art analyzers (1), quick-and-dirty methods are preferred (6, 13).

However, MRDs and corpora are not the only resources at hand. As Jacobs (2) is at pains to emphasize, the input text itself is a vital knowledge source that most NLP systems largely ignore. Exploiting this source implies a dynamic type of lexicon in which word senses are adapted or created during the processing of the input text. This approach, which represents a minority interest in this volume (12, 15), ties in well with the editor's main conclusion (1) that "systems must expressedly deal with lexical gaps as part of their normal operational mode. Satisfying this requirement entails a new lexical organizational principle, one that allows generalization, reasoning by analogy,

and lateral indexing." Martin (15) has defined explicit abstract structures with which metaphorical analogy can take place. Dyer (12) has implemented a connectionist model of "symbol recirculation" in which "words with similar semantics (as defined by word usage) end up forming similar distributed representations in the lexicon." Script-type word associations are acquired in similar fashion. Much of Dyer's argument finds echoes in my corpus-based approach to semantics (Sadler 1989).

NLP systems should be able to acquire knowledge of previously unknown words or usages from the (con)texts in which they are found. Thus, beside analogical capacities, systems need the ability to integrate the conclusions reached by analogy into their lexicon. Without such learning abilities there would appear to be no way out of the vicious circle: lexical acquisition depends on linguistic pre-processing of text, but effective pre-processing requires a comprehensive lexicon. And without a dynamic view of the lexicon, it is difficult to envision any efficient means of customizing it to different domains.

This book certainly provides a useful overview of the field, and there are some valuable pointers to the way ahead. There is a good deal of overlap (one paper has a half-page quote from another in the same volume), and relevance might have been improved by a tighter definition of the workshop's topic. To my mind there is one striking omission, and that is the enormous importance of bilingual sources. There is a passing reference to bilingual MRDs (10), but no mention whatever of bilingual corpora, in spite of the ongoing work at IBM (e.g., Dagan, Itai, and Schwall 1991; Brown et al. 1991) and BSO's Bilingual Knowledge Bank design (Sadler 1989). Many of the problems connected with recognition of word class and senses, idioms, and metaphors in text can be substantially resolved if such units are mapped onto equivalents in a different language.

Finally, one mean little grouse. Is it because computational linguists have so little faith in their own products that no one ran a spelling checker on this book?

References

Brown, Peter F.; Della Pietra, Stephen A.; Della Pietra, Vincent J.; and Mercer, Robert L. (1991). "Word-sense disambiguation using statistical methods." *Proceedings, 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, 264–270.

Dagan, Ido; Itai, Alon; and Schwall, Ulrike (1991). "Two languages are more informative than one." *Proceedings, 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, 130–137.

Sadler, Victor (1989). *Working with Analogical Semantics: Disambiguation Techniques in DLT*. Foris.

Victor Sadler is a senior linguistic systems designer at BSO/AI in Utrecht. Previously, he was responsible for lexicon development in the DLT machine translation project. Sadler's address is: BSO/AI, P.O. Box 8223, NL-3503 RE Utrecht, Netherlands.