Book Review

Corpus Processing for Lexical Acquisition

Branimir Boguraev and James Pustejovsky (editors)
(Apple Computer and Brandeis University)

Cambridge, MA: The MIT Press
(Language, speech, and communication
series), 1996, xvii+245 pp; hardbound,
ISBN 0-262-02392-X, $32.50

Reviewed by
Brian Ulicny
Inso Corporation

As the editors of this collection point out, NLP systems are generally only as good as the lexical resources that they employ. But acquiring these resources is costly. To what extent can the acquisition, extension, or improvement of computational lexical resources be automated? *Corpus Processing for Lexical Acquisition* contains revised versions of 10 papers originally presented at the ACL/SIGLEX Workshop on the Acquisition of Lexical Knowledge from Text, held at The Ohio State University in 1993, which focused on just these issues. The editors state that the contributions represent the "beginning of a new research programme combining linguistically inspired language analysis and statistically based corpus research" (p. 17). In addition to an introductory chapter by the editors providing a short review of the problems and previous approaches to lexical acquisition, the contributions are grouped into five sections: "Coping with unknown lexicalizations" (3 papers), "Task-driven lexicon induction" (2 papers), "Categorization of lexical units" (2 papers), "Lexical semantics from corpus analysis" (2 papers), and "Measuring lexical acquisition" (1 paper).

The first section, "Coping with unknown lexicalizations," deals primarily with identifying and reidentifying the names of individuals and organizations in journalistic text, particularly the *Wall Street Journal*. Largely due to the complications of their internal punctuation and irregular capitalization, naive approaches to name identification will not produce acceptable results. David D. McDonald's contribution "Internal and external evidence in the identification and semantic categorization of proper names" stands out as a state-of-the-art presentation of a context-sensitive grammar for proper-name identification. Here, internal evidence refers to the properties of the words of a sequence to be evaluated as a proper name, and external evidence refers to the properties of the context of that sequence. McDonald's system is especially interesting in that it does not rely on lists of open-class name elements, such as common first names and surnames. Inderjeet Mani and T. Richard MacMillan address similar issues in their contribution. Their paper, "Identifying unknown proper names in newswire text," outlines a sophisticated algorithm for identifying coreferential long names (*e.g., U.S. President Bill Clinton*) and short names (*e.g., Clinton*) within and across texts, drawing on discourse theory.

The general utility of this name recognition section is somewhat limited by its focus on the *Wall Street Journal* corpus. While McDonald reports impressive results for his system, it is fine-tuned to the particularities of the *Journal's* domain. For exam-

ple, McDonald's system takes proper names not to extend across possessives (p. 35), thereby excluding *Bram Stoker's Dracula* and *Babette's Feast* (two recent film titles) as single names, not to mention *sex, lies, and videotape* (all lowercase). For comparison, it would be helpful to see some discussion of name recognition and resolution in very poorly edited text, such as Usenet newsgroup corpora.

The second section, "Task-driven lexicon induction," reports work on the application of machine learning techniques to the problem of word-sense discrimination. Marti Hearst and Hinrich Schütze describe a variety of techniques for modifying the WordNet lexicon by means of lexical co-occurrence statistics from a specialized corpus in order to better identify the subject matter of documents. Similar techniques are proposed and evaluated in the contribution by Claudia Leacock and her colleagues Geoffrey Towell and Ellen Voorhees.

The third and fourth sections deal largely with the acquisition of argument structure from corpora and verb categorization. Roberto Basili and his colleagues report on their CIAULA system for inducing verb classifications within specialized text domains and provide an interesting discussion of the methodological considerations that went into its design. The system seems to require sophisticated manual markup of subcategorization patterns as input to the system. Basili et al. report that the verb clusters discovered may be "unintuitive" (p. 127), but it was unclear how the resulting clusters of verbs with varying adicities and subcategorization frames constituted a reasonable grouping of verbs. Here, perhaps, Beth Levin's work on English verb classes (Levin 1993) might serve as a better point of departure for future work in both English and other languages (cf. Jones et al. 1994).

Scott Waterman's contribution makes interesting use of the notion of "edit distance" in pattern matching, which has been successfully applied to problems in genetics, handwriting analysis, and other fields, in the problem of extracting the subcategorization patterns of various prepositions.

Victor Poznański and Antonio Sanfilippo give a crisp overview of their CorPSE system (Corpus-based Predicate Structure Extractor), which combines information from the *Longman Lexicon of Contemporary English* with corpus data in order to automatically assign thematic roles and verb senses in parsing. Somewhat less clear is Chinatsu Aone and Douglas McKee's contribution. They seek to automatically assign all verbs in English, Spanish, and Japanese to four **situation types**: *caused process*, *process-or-state*, *agentive-action*, and *inverse-state*. No linguistic motivation is provided for this very unintuitive set of classifications, so it is nearly impossible to evaluate their work. Aone and McKee, for example, inexplicably assert that *suffice* is correctly classified as a transitive *inverse-state* verb with a Goal subject and Theme direct object. As with the Basili contribution, if thematic roles are to be invoked, some account of how many there are and what they contribute to the sentence should be given.

The sole contribution to the section on evaluation is Gregory Grefenstette's paper proposing and comparing some methods for evaluating automatic assessments of word similarity. The automatic techniques are evaluated in comparison with machine-readable dictionaries and thesauri. The dictionaries and thesauri are proposed as "gold standards," but readers may make their own judgments as to which resource, the corpora or the lexicographic resources, get word similarities right.

In general, the editors' promise that this book represents a new marriage of linguistic and empirical techniques is somewhat overstated on the linguistic side. Little use is made of such notions as head of a phrase, scope, movement, inflection, dominance, and other basic concepts of linguistic analysis. While some attention is paid to passivization in the discussion of subcategorization frames, for example, there is no attempt to control for the effects of *wh*-movement in questions or relative clauses in

gathering a verb's syntactic frames. Also, as mentioned above, little attention is paid to how recent discussions of thematical roles and linking theory (most recently as in Dowty [1991], Hale and Keyser [1993], and Pesetsky [1995]) inform computational techniques, or how computational techniques could put them to the test.

Nevertheless, this collection is a valuable resource for those who need to address problems associated with the automatic acquisition of lexical resources, particularly with regard to proper-name identification and argument-structure assignment. It contains a thorough bibliography, and separate author and subject indices. No significant typos were found, although grammatical repairs were needed in some papers.

References

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67:547–619.

Hale, Kenneth and Samuel Jay Keyser. 1993. On argument structure and the lexical expression of syntactic relations. In Kenneth Hale and Samuel Jay Keyser, editors, *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, The MIT Press, Cambridge, MA.

Jones, Douglas A., Robert C. Berwick, Franklin Cho, Zeeshan R. Khan, Naoyuki Nomura, Anand Radhakrishnan, Ulrich Sauerland, and Brian Ulicny. 1994. Verb classes and alternations in Bangla, German, English, and Korean. Memo 1517, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago, IL.

Pesetsky, David. 1995. *Zero Syntax: Experiencers and Cascades.* The MIT Press, Cambridge, MA.

Brian Ulicny completed his doctorate at MIT in 1993 and is a Senior Member of the Technical Staff at Quarterdeck Corporation. His research interests include the application of lexical semantics to information retrieval tasks. Ulicny's address is: Inso Corporation, 31 St. James Avenue, Boston, MA 02116; e-mail: bulicny@inso.com