Challenges in Natural Language Processing

Madeleine Bates and Ralph M. Weischedel (editors)
(BBN Systems and Technologies)

Cambridge, England: Cambridge
University Press (Studies in natural
language processing, edited by
Branimir K. Boguraev), 1993, xi +
296 pp.
Hardbound, ISBN 0-521-41015-0, $49.95

Reviewed by
Eduard Hovy
Information Sciences Institute, USC

Prognostication is a dangerous business. Doing it in print is even more dangerous. Unless your predictions are luckily, spectacularly right, people in the future will read your milder (but correct) predictions with a yawn and, worse, your failures with amusement. The way to protect your dignity, of course, is not to make real predictions, but merely to discuss trends and to focus on issues that are unlikely to be decided for a long time.

That is what happens in this book. A collection of papers from a workshop held at Bolt Beranek and Newman, Inc. (BBN) in Cambridge, Massachusetts, at the end of 1989, the book records the predictions, expectations, and trend analyses of a number of prominent researchers in the areas of computational linguistics, lexicography, phonology, and related areas. As explained in the preface, the purpose of the workshop was to "discuss the most significant challenges and problems that will face the field of computational linguistics in the next two to ten years." For those who care about the forces that largely shape developments in our field (i.e., politics and money), it will be evident that a second (if not first) purpose of the workshop was to argue before funding agents from (then) DARPA and (then) RADC that computational linguistics had not come to a relative standstill in the 1980s, but was still going strong, held significant promise, and was going to make great strides over the next two to ten years. That was five years ago. We are in an excellent position to judge the prognostications of the authors.

In its overt purpose, the book does not succeed. In one way or another, the authors all show their caution and wisdom in avoiding anything other than general predictions and in sticking to the tried-and-true formula of describing hard problems and hinting at promising directions for solutions. But because the unsolved hard problems discussed in the papers remain unsolved hard problems five years later, they are by and large worth reading. However, the workshop did not fail in its covert purpose, since US funding for natural language processing remained strong over the past five years, a fact for which everyone in the field possibly owes the workshop organizers (the book editors) a debt of gratitude.

The book is organized into six parts, each devoted to a theme: challenging problems, the lexicon, semantics and knowledge representation, discourse, speech, and in conclusion, a list of problems for the near future.

Part I contains only one paper, by Madeleine Bates, Rusty Bobrow, and Ralph Weischedel (the workshop home team from BBN) and is a survey of written language processing from the BBN perspective. Mostly, the paper provides lists: the major ar-

eas of computational linguistics as syntax, semantics, and pragmatics, useful types of knowledge acquisition, types of incompleteness in input language, and open research problems. The heart of the paper is a fairly detailed description of aspects of BBN's written language-processing systems and representations. Overall, this paper can be read as a snapshot of one approach to natural language processing circa 1989.

Part II of the book is devoted to the lexicon and contains three papers of more general interest. Sue Atkins opens with a delightful paper describing just how little dictionaries can be trusted, by comparing various dictionaries' entries for word clusters such as {*safety, danger, risk*} and {*admire, acknowledge, admit*}. She identifies core meaning components in order to motivate the structure of an idealized machine-readable dictionary, underscoring that the problems inherent in lexicography were apparent from its very inception by ending with a 1755 quotation from Samuel Johnson himself:

> [K]indred senses may be so interwoven, that the perplexity cannot be disentangled, nor any reason be assigned why one should be ranged before the other. When the radical idea branches out into parallel ramifications, how can a consecutive series be formed of senses in their nature collateral? The shades of meaning sometimes pass imperceptibly into each other, so that though on one side they apparently differ, yet it is impossible to mark the point of contact. Ideas of the same race, though not exactly alike, are sometimes so little different, that no words can express the dissimilitude, though the mind easily perceives it, when they are exhibited together; and sometimes there is such a confusion of acceptations, that discernment is wearied, and distinction puzzled, and perseverance herself hurries to an end, by crowding together what she cannot separate.

Beth Levin's paper follows smoothly by describing what (and how little) linguistics can contribute to the lexicon effort, even in the case of seemingly simple verbs of sound like {*whistle, grunt, bleep, bellow*}, by bringing to bear linguistic argumentation about syntactic properties, selectional restrictions, collocations, etc.

Bran Boguraev's paper complements the others by describing problems inherent in extracting lexical information from machine-readable dictionaries. Only with a prior theory to guide and structure the results of the extraction effort, Boguraev argues, is one able to find and appreciate nonsuperficial information, whether by "trawling" through machine-readable dictionaries or by minutely examining entries. In either case, he says, building a lexicon for computational purposes from machine-readable dictionaries is never a process of "'cranking the handle' and getting a lexicon overnight," but rather of carefully designing a lexicon and then, for each aspect of lexical data, carefully searching each source in its entirety for useful information of any form. He provides examples of internal dictionary representations and the kinds of processing required to work with them.

Together, these three papers show just how hard lexical semantics is, how little we know about it to date, and how hard it is going to be to build up adequate computational lexicons. The papers should be required reading for all students of computational linguistics and knowledge representation.

Having experienced a proper sense of humility, we move on to Part III, the section on semantics and knowledge representation. Robert Moore's paper identifies a problem in the treatment of events in the accounts of both Davidson and Perry (and Barwise). Studying adverbials, he notes the difference between *John sang strangely* and

Strangely, John sang: in the first, the manner of John's singing was strange, whereas in the second, the fact that John sang was strange. To handle this representationally, one clearly needs two entities for *strange* to modify: one for the actual singing, and one for the event of the singing. Moore adopts Perry's approach to associating situations with whole sentences, because this supplies the two requisite entities in a convenient form, but he adopts Davidson's approach to representing adverbs.

The other paper in this section, by James Allen, discusses the distressing separation between parsing and semantic analysis (what Allen calls structural processing) on the one hand and knowledge representation (KR) and reasoning in current NL systems on the other. Despite early NL work that viewed these two aspects as inextricable, this separation is almost ubiquitous today, even in large, longer-lifespan systems that perform both structural processing and KR reasoning. Taking ambiguity as his lodestone, Allen provides numerous ways in which no system can be complete without supporting both kinds of processing and shows why performing structural processing and then reasoning in strict sequentiality is not only wasteful, but can be prohibitively expensive.

He then argues for three important points: The KR must support long-term representation of ambiguity (which involves distinguishing between ambiguity and disjunction, performing inference over ambiguity, and supporting disambiguation techniques); the KR cannot avoid having full expressive power; and compositional semantic interpretation need not be constrained by the final KR language. Throughout the paper, explicitly and implicitly, Allen hints at the kind of solution that is slowly coming to the fore nowadays and that I believe provides our only hope for robust yet deep NL understanding: incomplete and probabilistic reasoning. This is a long paper, but an important one to be read (and occasionally re-read) by everyone working on NL understanding and KR systems as a reminder of how much more there is to be done.

Part IV contains papers on discourse by Passonneau and Steedman. Passonneau's paper addresses the nature of the attentional state in a discourse—roughly, that set of entities upon which the interlocutors are concentrating at any point—by contrasting the operation of the pronouns *it* and *that*. Whereas *it* establishes what is called a local center (a reference point for future utterances), *that* changes the attentional state of an existing discourse entity or creates a new one. To prove her case, Passonneau counts the number of times in a corpus of natural dialogue that *it* and *that* were preceded by various syntactic entities—noun phrases (both subject and non-subject), pronouns, etc. In nine cases she finds statistically significant predictiveness from a class of syntactic entity to one of the two pronouns. Although the paper is exhaustive in its discussion of centering and attentional states (and it points out problems as well as strengths of these notions), it suffers from including hardly anything at all about the data, such as how much there was, where it came from, and so on.

Steedman's paper addresses the problem that sentence intonation structure and sentence syntactic structure in most accounts of syntax do not line up. For many people this is not a problem, of course—Halliday would simply provide the two analyses in parallel—but Steedman argues that with his combinatory categorial grammar (CCG), one can achieve isomorphism essentially by creating intonational categories whose combination rules constrain the syntactic combinations (of which there are far more in CCG than in other grammars). That is, a CCG analysis of a sentence on the purely syntactic level would support many readings with odd-looking (to a syntactician of another school) intermediate nodes and groupings; however, argues Steedman, these categories may be real units when you take intonation into account—for example, when the sentence under analysis answers a question. Steedman uses Pierrehumbert's

notation involving pitch accents and boundaries. (One problem with the book is that Pierrehumbert's paper follows Steedman's, requiring the uninformed reader to skip around.)

Part V comprises Pierrehumbert's paper on prosody and intonation in speech technology. As is well known, these two phenomena have important effects on meaning, as, for example, in the difference in truth value of these two statements (where uppercase indicates stress):

> In English, a *U* usually follows a *q*.
> In English, a *u* usually follows a *Q*.

The first is true and the second is false. Pierrehumbert describes why the current dominant speech-processing technology, hidden Markov models, is inadequate for capturing these phenomena; what is more, phoneticians have not yet succeeded in creating a comprehensive quantitative model of the sound structure for even one voice. She then discusses various aspects of a model of sound structure and its use, including nonlocal effects, focus, and tunes. As with the previous two papers, this paper is best appreciated by a reader with some prior background in the area.

Part VI is a brief listing of critical areas for work in natural language processing (where "critical" means impact on technology) and the attendant support resources and organizations required, by the editors of the book. This chapter is best appreciated by a potential funder of NLP research.

Eduard Hovy is one of the principal investigators of the Pangloss machine translation project and has also worked on discourse structure and multimedia presentation management. Hovy's address is USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292-6695. E-mail: hovy@isi.edu