Briefly Noted

Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology

Maria Teresa Pazienza (editor)
(University of Rome, Tor Vergata)

Berlin: Springer-Verlag (Lecture notes in computer science, volume 1299), 1997, ix+213 pp; paperbound, ISBN 3-540-63438-X, $43.00

This book is a diverse collection of ten presentations given at an International Summer School on Information Extraction in Rome, 1997. The goal of information extraction (IE) is selective, task-driven interpretation of text narrative in order to fill out templates with information about a particular scenario. I was disappointed to find that only five of the articles were actually about IE research. The other half of the articles addressed issues peripheral to IE, such as information retrieval (IR) and text classification.

The first two articles are by Yorick Wilks and by Ralph Grishman, who are prominent IE researchers in the UK and the US, respectively. Each gives a high-level discussion of IE, its successes and limitations. Wilks makes the observation that IE's strength comes from its modular architecture. Individual modules such as part-of-speech tagging or morphology analysis can be constructed and optimized independently and reused in a variety of applications. He sees the primary limitation of IE to be the template representation that restricts the type of information that can be extracted.

Grishman describes the typical architecture of IE systems whose modules include lexical analysis, name recognition, shallow syntactic parsing, task-specific pattern matching, coreference analysis, event merging, and finally template generation. He identifies the main challenges to IE as the cost of adapting a system to a new domain or scenario and a ceiling on performance, which is closely related to the issue of knowledge acquisition and difficulty handling complex syntactic structures.

Three other articles deal with more specialized topics within IE. Robert Gaizauskas, Kevin Humphreys, Saliha Azzam, and Yorick Wilks describe a system for multilingual IE with some language-independent modules that are indexed by language-specific lexicons. Roberto Basili and Maria Teresa Pazienza discuss corpus-driven lexical acquisition, in particular for the "foreground" lexicon of words that support a particular IE task. Branimir Boguraev and Christopher Kennedy present work in technical-term recognition and how this can be a step towards document summarization.

The remaining articles concern IR, text classification, or heterogeneous database techniques, and are only tangentially related to IE. Gregory Grefenstette presents an NLP-based strategy for suggesting additional IR query terms to a user. Alan Smeaton gives a tutorial on uses of NLP in IR. Nicola Guarino discusses formal ontologies and how these can enhance IR with semantic matching. Filippo Neri and Lorenza Saitta give a tutorial on machine learning that briefly touches on text classification. Sophie Cluet describes database techniques for querying semi-structured Web pages.—*Stephen Soderland, Children's Hospital, Seattle*

Observing Interaction: An Introduction to Sequential Analysis (second edition)

Roger Bakeman and John M. Gottman
(Georgia State University and University of Washington)

Cambridge, England: Cambridge University Press, 1997, 207 pp; hardbound, ISBN 0-521-45008-X, £50.00, $69.95; paperbound, ISBN 0-521-57427-7, £17.95, $24.95

The current trend for "empirical methods" in computational linguistics (see, for example, the special issue of *Computational Linguistics*, 23(1), March 1997) has led many researchers to mark up dialogue and text corpora subjectively for linguistic phenomena such as discourse structure and coreference. This has led to some methodological confusion, especially since as a field computational linguistics has not used these techniques before. However, the idea of "coding" is not new; within psychology there is a long tradition of systematically observing behavior so that statistical analysis can be performed. Even something akin to dialogue-move classification for group discussion was developed in the 1950s (Bales 1950) and has been used in various guises ever since. Bakeman and Gottman's *Observing Interaction: An Introduction to Sequential Analysis*, recently out

in a second edition, is an excellent manual for observational analysis techniques. It contains advice about developing coding systems, testing agreement among coders, and analyzing the results. Although many different kinds of behaviors are included (notably the state of children's play and sounds made by baby chicks before they hatch), many of the examples are based on Gottman's work on marital conversations, and so relate fairly closely to research in this community.

This manual is not a perfect fit for computational linguists because in psychology, coding is designed to answer a specific research question and so it is fairly simple. The authors remind the reader that coding schemes should be tailored for the research question under study and not borrowed. This makes for much better hypothesis testing—and we would do well to remember that—but it will grate with people developing coding schemes in order to train language engines, since there the considerations are rather different. Psychologists do not often use more than one or two flat-structured behavioral codings at a time, so there is not much advice about relating different codes or comparing the distribution of codes in data collected under two different conditions. To save effort, coding of linguistic behavior is often performed directly from videotape rather than resorting to transcripts, and so there are fewer representational issues involved. Often results are expressed simply in terms of differences in code counts, although the point of this book and the companion volume describing software tools (Bakeman and Quera 1992) is to convince observational analysts that it's possible to perform much more sophisticated analyses that describe patterns in how behaviors are sequenced. The book also assumes knowledge of very basic inferential statistics but describes transitional probabilities at great length, which is exactly the wrong way round for most of us.

Despite these difficulties, I highly recommend this book to all researchers setting out on coding exercises. The coding advice given is always insightful, often lively, and addresses many of the concerns that empirical researchers are discussing behind the scenes. Even if there were a manual tailored for our community, I still think there would be merit in reading this one. Coding-based research is much more mature in psychology than in computational linguistics, and although for practical reasons we may not wish to emulate the studies described, just thinking about the issues involved is likely to improve our work. In addition, the methods for sequential analysis, plus a distinction Bakeman and Gottman make between microcoding and macrocoding, which codes for sequences of events, are potentially quite useful for informing systems that need good predictions of behavior, like dialogue engines. One word of warning, though: although the coding advice has not changed very much between the editions, in the chapters about analysis the second edition corrects a basic statistical mistake and adds a whole new and much better technique based on log-linear modeling. Glance through the library's first edition if you will, but if you want to use the analytical methods, you will have to order the new one.—*Jean Carletta, University of Edinburgh*

References

Bakeman, Roger, and Vincenç Quera. 1992. *Analyzing Interaction: Sequential Analysis with SDIS and GSEQ.* Cambridge University Press.

Bales, Robert Freed. 1950. *Interaction Process Analysis: A Method for the Study of Small Groups.* Addison-Wesley Press, Cambridge, MA.

Representation and Processing of Spatial Expressions

Patrick Olivier and Klaus-Peter Gapp (editors)
(University of Wales, Aberystwyth and Universität des Saarlandes)

Mahwah, NJ: Lawrence Erlbaum Associates, 1998, viii+287 pp; hardbound, ISBN 0-8058-2285-2, $79.95 (special prepaid price, $49.95)

This book presents 16 new papers on the linguistic expression of spatial relations. This general topic has attracted a good deal of attention recently. It raises fundamental questions about the linking of word to world, and can be profitably approached from the perspectives of several disciplines. The chapters of this book reflect the multidisciplinary na-

ture of the subject matter, presenting both computational and psycholinguistic studies of this topic. In this review, I briefly discuss a selection of the chapters, some from each of these two disciplines.

The computational chapters provide an overview of some of the challenges that arise in the processing of spatial expressions, and the techniques employed to address them. A central issue addressed by these chapters is that of appropriate knowledge representation. Fuhr, Socher, Scheering, and Sagerer present a clear and reasonable-seeming approach in which the space surrounding a reference object is subdivided into fairly fine regions, and the applicability of a spatial term is a function of the region or regions into which the located object falls. In an interesting variation on this region-based theme, Edwards and Moulin present a system that uses Voronoï diagrams to recognize and classify spatial relations among objects. In contrast to these region-based and essentially geometric approaches, Di Tomaso, Lombardo, and Lesmo argue convincingly for the use of semantic networks of detailed world knowledge in the processing of some spatial expressions, and present a system based on this principle. Blocher and Stopp address a different problem, that of producing an appropriate spatial utterance given limited time to do so. They present an anytime-algorithm for this purpose, which returns descriptions of increasing accuracy given increasing amounts of processing time. And the chapter by Voss, Dorr, and Sencan discusses the problem of machine translation of spatial expressions from English to Turkish. These two languages differ as to when a spatial verb particle obligatorily accompanies a verb. In Voss, Dorr, and Sencan's interlingua-based system, this problem is addressed through semantic markings in the lexicon.

The psycholinguistic chapters provide a useful overview of some recent empirical work bearing on the processing of spatial language. In the most linguistically oriented of these chapters, Herskovits argues that the role of *schematization* in spatial language is more subtle than has so far been appreciated. Other chapters report on experimental work. Spivey-Knowlton, Tanenhaus, Eberhard, and Sedivy demonstrate compellingly that visuospatial context can influence the on-line processing of spatial language—and thus, that syntactic processing is not informationally encapsulated. The chapter by Bryant provides useful coverage of some recent experimental work on the relative psychological salience of the different body axes in spatial cognition and language. Schober's chapter demonstrates that a speaker's spatial utterances often give evidence of an understanding of the addressee's viewpoint, which might or might not be the same as that of the speaker. And Coventry's chapter presents experimental evidence implicating functional, rather than purely geometric, knowledge in spatial language.

While the book treats both computational and psycholinguistic work, it does not in general bring these two streams of inquiry into very close contact with one another. (An exception is Di Tomaso, Lombardo, and Lesmo's and Coventry's shared emphasis on detailed knowledge of object function, as a critical element of spatial language.) The reader will also find that the chapters are of uneven quality. However, some are excellent, and will be useful to those wishing to learn about current research in spatial language.
—Terry Regier, University of Chicago

Computing Natural Language

Atocha Aliseda, Rob van Glabbeek, and Dag Westerståhl (editors)
(Stanford University and University of Stockholm)

Stanford: CSLI Publications (CSLI lecture notes, number 81) (distributed by Cambridge University Press), 1998, x+158 pp; hardbound, ISBN 1-57586-101-1, $59.95; paperbound, ISBN 1-57586-100-3, $22.95

The volume is an "outgrowth" of the Fourth CSLI Workshop on Logic, Language, and Computation (Stanford, 2–4 June 1995). The contents of the volume are as follows:

"Indexicals, contexts and unarticulated constituents" by John Perry.
"Formalizing context (Expanded notes)" by John McCarthy and Saša Buvač.
"Changing contexts and shifting assertions" by Johan van Benthem.
"Discourse preferences in dynamic logic" by Jan Jaspars and Megumi Kameyama.

"Polarity, predicates and monotonicity" by Víctor Sánchez Valencia.
"HPSG as type theory" by M. Andrew Moshier.
"Machine learning of physics word problems: A preliminary report" by Patrick Suppes, Michael Böttner, and Lin Liang.

Phonological Representations: Their Names, Forms and Powers

John Coleman
(Oxford University)

Cambridge University Press (Cambridge studies in linguistics, volume 85), 1998, xvii+345 pp; hardbound, ISBN 0-521-47208-3, $74.95

"Rewriting rules, derivations, and underlying representations are enduring characteristics of generative phonology. In this book, John Coleman argues that they are unnecessary. The expressive resources of context-free unification grammars are sufficient to characterize phonological structures and alternations.

"According to this view, all phonological forms and constraints are partial descriptions of surface representations. This framework, now called Declarative Phonology, is based on a detailed examination of the formalisms of feature theory, syllable theory, and the leading varieties of non-linear phonology. Dr. Coleman illustrates this with two extensive analyses of the phonological structure of words in English and Japanese. As Declarative Phonology is surface-based and highly restrictive, it is consistent with findings in cognitive psychology and amenable to straightforward computational implementation."—*From the publisher's announcement*