Briefly Noted

Towards High-Precision Machine Translation: Based on Contrastive Textology

John Laffling

Berlin: Foris Publications (Distributed Language Translation Series, edited by Toon Witkam), 1991, xiii + 178 pp. Hardbound, ISBN 3-11-013388-1, DM 98.00

High-Precision Machine Translation is defined in this book as machine translation for dissemination, with particular emphasis on choosing a target-language term that not only adequately renders the source language term, but also is the going term in the particular sub-language of that text. The high-level goal is to produce translations that read as fluently and as competently as text on the same topic written in the target language. The introduction argues for separating linguistic from world knowledge, claiming that linguistic knowledge alone can yield full-blown semantic modeling, at least within the somewhat limited domain of a sub-language. The author claims that semantic modeling is required for machine translation and sets his sub-language to be *party political programs* in German. This sub-language, he argues, is rich enough to be more than a toy domain but stylized enough to limit the size of the dictionary to be developed.

Chapter 2 details the importance of the co-text for disambiguation and the important role of coherence and cohesion in the reception of text, thought to be embodied in a thesaurus of synonyms. Chapter 3 then tears into the "sentence-based" approach of most traditional machine translation systems because it leads to a statistical determination of the equivalence of isolated signs in two languages and cannot yield pragmatic equivalence. Chapter 4 introduces some of the technical contributions of the book. A dictionary of thesaural relations of 200 nouns is compiled for the selected "corpus" of party programs of 60,000 words. The stylistically pure corpus is intended to yield stylistically adequate thesaural relations, i.e., pragmatically equivalent translations. Disambiguation proceeds by weighting and counting co-occurrence of the noun with any of the synonyms, near synonyms, or significantly co-occurring adjectives of the analysis dictionary that occur within a window which varies from four words to seven sentences, depending on the phase of the attempted disambiguation process. Three phases with increasingly relaxed constraints on the matching procedure are possible; the process stops when disambiguation is achieved.

A separate transfer dictionary is described in Chapter 5, again of the form of thesaural relations paired with translation equivalents, now for the noun in question when occurring in the environment of one or more of the contextually relevant terms. These translation equivalents are found by poring over "parallel texts," by which Laffling means texts of the same sub-language and style that address similar points. Chapter 6 summarizes the effort as based on linguistic principles and points out several shortcomings, such as not testing the system on texts that were not part of the corpus from which the dictionaries were compiled, differences in performance when dealing with (wordy) party programs (good performance) and (concise) party manifestos (worse performance), and lack of any attempt to put together the proposed translation equivalents to actually produce a target-language text.

The book presents these ideas awkwardly hidden in a text that violates all the expectations I have for a computational linguistics volume. The algorithms and procedures are not declared explicitly; the data are not described in detail; no attempt at evaluation of the procedures or the resulting dictionaries is made; the discussion of the literature focuses on old work (Schank's original scripts are discussed at length) and tends to set up straw men for easy rebuttal. The tone is at times unnecessarily polarizing. But most aggravating is the style of presentation. The author does not state his points clearly and gives no data-supported arguments for and against his claims. Instead he selects disembodied quotes from a wide range of authors to stand in for what he wants to say. These quotes are frequently in German, sometimes in French, without any translation. The work of the quoted authors is not summarized and does not always lend credence to the point Laffling has in mind. (Most amusing to me was the use of a quote by Pustejovsky to exemplify how "R. Schank and others involved on the Yale Project argue that connectivity is only implicit in a text..." [p. 48].)

The system is implemented on a PC in Basic, and much of the (little) technical discussion centers on working around these artificial limitations. The presented dictionaries appear to have been compiled by hand from some corpus analysis (concordances are hinted at) and (arbitrarily?) refined until they perform the task of disambiguating occurrences in the corpus and suggesting the most suitable translation for that word (or phrase).

The 22-page bibliography (over 10% of the length of the book!) was out of date even at the date of publication (1991), the newest reference being from 1989. But with its focus on older and European publications, it does provide a complementary set of references that are often unavailable in North American libraries (and publications). The usefulness of this extensive bibliography is, however, lessened by the fact that the text cannot be trusted to serve as an annotation to it.

Overall, the book leaves the reader with the uneasy feeling of having missed out on a thorough introduction into interesting research and having been forced to read through a thesis.—*Sabine Bergler, Concordia University*

Current Issues in Computational Linguistics: In Honour of Don Walker

Antonio Zampolli, Nicoletta Calzolari, and Martha Palmer (editors)
(Universitá di Pisa; Istituto di Linguistica Computationale del CNR, Pisa; and the University of Pennsylvania)

Dordrecht: Kluwer Academic Publishers,
1994, and Pisa: Giardini editori e stampatori
in Pisa (Linguistica Computazionale,
volume IX–X), 1994, xxv + 594 pp.
Hardbound, ISBN 0-7923-2997-X, $74.00;
Paperbound, ISBN 0-7923-2998-8, $39.00[1]

"This collection of papers is dedicated to Don Walker, a kind, gentle soul, and a man of intelligence and vision whom we have been privileged to know and to work with. The technical content of these papers clearly reflects his guidance and direction: his understanding of the relevance of theories from related disciplines such as linguistics, psychology, and philosophy; and his dream of the future availability of electronic documents and the potential of this as a resource for corpus-based analysis. This led to one of his most interdisciplinary roles, as a primary mover in the push to standardize the acquisition and tagging of electronic text, an endeavor that brought together people from fields as distinct as publishing, the humanities and computer science. His wide-spread interests are mirrored in the rich and diverse collection of papers in this volume which portray many different approaches and many different styles. These papers co-exist in harmony here, united by the goal of paying tribute to Don as well as actualizing part of this dream."—*From the introduction by Martha Palmer*

The contents of the volume are the following. Those marked with an asterisk are original to the volume; the other papers are revisions or reprints of work previously published or presented.

"Donald Walker: A remembrance"

Section 1: The task of natural language processing

"Natural language processing: A historical review" by Karen Sparck Jones;

"On getting a computer to listen" by Jane Robinson;

"Utterance and objective: Issues in natural language communication" by Barbara Grosz;

"On the proper place of semantics in machine translation" by Margaret King.

Section 2: Building computational lexicons

"Developing a natural language interface to complex data" by Gary G. Hendrix, Earl D. Sacerdoti, Daniel Sagalowicz, and Jonathan Slocum;

"User-needs analysis and design methodology for an automated document generator" by Karen Kukich, Kathleen McKeown, J. Shaw, J. Robin, J. Lim, N. Morgan, and J. Phillips;

"Machine-readable dictionaries and computational linguistics research" by Branimir Boguraev;

"Research toward the development of a lexical knowledge base for natural language processing" by Robert A. Amsler;

*"Discovering relationships among word senses" by Roy J. Byrd;

*"Machine readable dictionary as a source of grammatical information" by Eva Hajičová and Alexandr Rosen;

[1] Royalties from the volume will be donated to the ACL's Don and Betty Walker International Student Fund.

*"The IIT lexical database: Dream and reality" by Sumali Pin-Ngern Conlon, Joanne Dardaine, Agnes D'Souza, Martha Evens, Sherwood Haynes, Jong-Sun Kim, and Robert Strutz;
*"Visions of the digital library: Views on using computational linguistics and semantic nets in information retrieval" by Judith L. Klavans;
"Anatomy of a verb entry: From linguistic theory to lexicographic practice" by Beryl T. Atkins, Judy Kegl, and Beth Levin;
"Issues for lexicon building" by Nicoletta Calzolari;
"Outline of a model for lexical databases" by Nancy Ide, Jacques Le Maitre, and Jean Véronis;
*"Construction-based MT lexicons" by Lori Levin and Sergei Nirenburg;
*"Dependency-based grammatical information in the lexicon" by Petr Sgall;
*"Semantics in the brain's lexicon—Some preliminary remarks on its epistemology" by Helmut Schnelle.

Section 3: The acquisition and use of large corpora
"The ecology of language" by Donald E. Walker;
"Representativeness in corpus design" by Douglas Biber;
*"The Text Encoding Initiative" by C. M. Sperberg-McQueen;
*"Discrimination decisions for 100,000-dimensional spaces" by William A. Gale, Kenneth W. Church, and David Yarowsky;
"Acquisition and exploitation of textural resources for NLP" by Susan Armstrong-Warwick;
*"The Center for Electronic Texts in the Humanities" by Susan Hockey;
"Design principles for electronic textual resources: Investigating users and uses of scholarly information" by Nicholas J. Belkin.

Section 4: Topics, methods and formalisms in syntax, semantics and pragmatics
"Some recent trends in natural language processing" by Aravind K. Joshi;
"Two principles of parse preference" by Jerry R. Hobbs and John Bear;
"Varieties of heuristics in sentence parsing" by Makoto Nagao;
"UD: Yet another unification device" by R. Johnson and M. Rosner;
"Evaluating English sentences in a logical model" by Joyce Friedman, Douglas B. Moran, and David Scott Warren;
"Recovering implicit information" by Martha S. Palmer, Deborah A. Dahl, Rebecca J. Schiffman, Lynette Hirschman, Marcia Linebarger, and John Dowding;
*"Flexible generation: Taking the user into account" by Cécile L. Paris and Vibhu O. Mittal;
"Stone soup and the French room" by Yorick Wilks.

Corpus-Based Research into Language: In Honour of Jan Aarts

Nelleke Oostdijk and Pieter de Haan (editors)
(University of Nijmegen)

Amsterdam: Editions Rodopi (Language and Computers: Studies in Practical Linguistics, edited by Jan Aarts and Willem Meijs, volume 12), 1994, vii + 279 pp.
Paperbound, ISBN 90-5183-588-4, $50.00, Dfl 85.00

"For over two decades Jan Aarts has been actively involved in corpus linguistic research. He was the instigator of a large number of projects, and he was responsible for what has become known as 'the Nijmegen approach' to corpus linguistics. ... The present volume has been collected in his honour."—*From the publisher's announcement*

The contents of the volume are the following:

"A tribute to Jan Aarts" by Flor Aarts;
"Introduction" by Nelleke Oostdijk and Pieter de Haan;
"Continuity and change in the encoding of computer corpora" by Stig Johansson;
"Tagging the British ICE Corpus: English word classes" by Sidney Greenbaum and Ni Yibin;
"The large-scale grammatical tagging of text: Experience with the British National Corpus" by Geoffrey Leech, Roger Garside, and Michael Bryant;
"Computerized lexicons and theoretical models" by Willem Meijs;
"Resolving lexical ambiguity" by Louise Guthrie, Joe Guthrie, and Jim Cowie;
"Prospects for practical parsing of unrestricted text: Robust statistical parsing

techniques" by Ted Briscoe;
"Robust parsing of unconstrained text" by Fred Karlsson;
"Using parsed corpora: A review of current practice" by Clive Souter and Eric Atwell;
"An experiment in customizing the Lancaster Treebank" by Ezra Black;
"SUSANNE: A Domesday Book of English grammar" by Geoffrey Sampson;
"What is wrong with adding one?" by William Gale and Kenneth Church;
"Intra-textual variation within medical research articles" by Douglas Biber and Edward Finegan;
"On the functions of *such* in spoken and written English" by Bengt Altenberg;
"Imparsable speech: Repeats and other nonfluencies in spoken English" by Anna-Brita Stenström and Jan Svartvik.

Natural Language Processing

Fernando C. N. Pereira and Barbara J. Grosz
(AT&T Bell Laboratories and Harvard University)

Cambridge, MA: The MIT Press, 1994,
531 pp.
Paperbound, ISBN 0-262-66092-X, $35.00
First published as a special issue of *Artificial Intelligence*, 63(1–2), 1993.

"This special issue aims to present to the general artificial intelligence community accounts of the major technical ideas underlying many of the significant advances in natural-language processing (NLP) over the last decade. The issue focuses in particular on those areas for which the main challenges are instances of general AI problems; thus the papers exhibit the strong connections between NLP and such other areas of AI research as knowledge representation, reasoning, planning, and integration of multiple knowledge sources."—*From the editors' introduction*

The contents of the volume are the following:

"Introduction" by Fernando C. N. Pereira and Barbara J. Grosz;
"The KERNEL text understanding system" by Martha S. Palmer, Rebecca J. Passonneau, Carl Weir, and Tim Finin;
"Interpretation as abduction" by Jerry R. Hobbs, Mark E. Stickel, Douglas E. Appelt, and Paul Martin;
"Innovations in text interpretation" by Paul S. Jacobs and Lisa F. Rau;
"Lexical knowledge representation and natural language processing" by James Pustejovsky and Branimir Boguraev;
"Parsing as non-Horn deduction" by Edward P. Stabler Jr.;
"Time and modality in a natural language interface to a planning system" by R. S. Crouch and S. G. Pulman;
"Pitch accent in context: Predicting intonational prominence from text" by Julia Hirschberg;
"Automated discourse generation using discourse structure relations" by Eduard H. Hovy;
"Plan-based integration of natural language and graphics generation" by Wolfgang Wahlster, Elisabeth André, Wolfgang Finkler, Hans-Jürgen Profitlich, and Thomas Rist;
"Interlingual machine translation: A parameterized approach" by Bonnie J. Dorr;
Review of *Meaning and Grammar: An Introduction to Semantics* (by Gennaro Chierchia and Sally McConnell-Ginet) by C. Raymond Perrault;
Review of *Language in Action: Categories, Lambdas and Dynamic Logic* (by Johan van Benthem) by David Israel;
Review of *Intentions in Communication* (edited by Philip R. Cohen, Jerry Morgan and Martha E. Pollack) by Jon Oberlander.

Infolingua

Conrad F. Sabourin (editor)
(Unisys Corporation; School District 833, St Paul Park, MN; Unisys Corporation; and University of Adelaide)

Montreal: Infolingua Inc, 1994.
ISSN 1198–1083

Infolingua is a series of 17 bibliographies on topics in computational linguistics and related areas of artificial intelligence, linguistics, informatics, communications, and education. The volumes in the series are listed below. The number of references in each is five to six times the number of pages.

More information is available from the publisher, Infolingua Inc., P.O. Box 187, Snowdon, Montreal, Canada H3X 3T4. E-mail *73651.2144@compuserve.com*

1. "Computational morphology," by Conrad F. Sabourin; 492 pp. ISBN 2-921173-01-8, $80.00
2. "Computational parsing," by Conrad F. Sabourin; 2 volumes, 1029 pp. ISBN 2-921173-02-6 and 2-921173-03-4, $150.00
3. "Computational lexicology and lexicography," by Conrad F. Sabourin; 2 volumes, 1031 pp. ISBN 2-921173-04-2 and 2-921173-05-0, $150.00
4. "Computational text understanding," by Conrad F. Sabourin; 657 pp. ISBN 2-921173-06-9, $80.00
5. "Computational text generation," by Conrad F. Sabourin; with a survey article by Mark T. Maybury; 649 pp. ISBN 2-921173-07-7, $80.00
6. "Natural language interfaces," by Conrad F. Sabourin; 2 volumes, 847 pp. ISBN 2-921173-08-5 and 2-921173-09-3, $130.00
7. "Machine translation," by Conrad F. Sabourin and Laurent R. Bourbeau; 2 volumes, 1168 pp. ISBN 2-921173-10-7 and 2-921173-11-5, $180.00
8. "Literary computing," by Conrad F. Sabourin; 581 pp. ISBN 2-921173-12-3, $80.00
9. "Computer-assisted language teaching," by Conrad F. Sabourin and Elca Tarrab; 2 volumes, 1066 pp. ISBN 2-921173-13-1 and 2-921173-14-X, $150.00
10. "Computer-mediated communication," by Conrad F. Sabourin and Rolande M. Lamarche; 2 volumes, 862 pp. ISBN 2-921173-15-8 and 2-921173-16-6, $130.00
11. "Electronic document processing," by Conrad F. Sabourin and Rolande M. Lamarche; 551 pp. ISBN 2-921173-17-4, $80.00
12. "Computational character processing," by Conrad F. Sabourin; 580 pp. ISBN 2-921173-18-2, $80.00
13. "Quantitative and statistical linguistics," by Conrad F. Sabourin; 508 pp. ISBN 2-921173-19-0, $80.00
14. "Mathematical and formal linguistics," by Conrad F. Sabourin; 612 pp. ISBN 2-921173-20-4, $80.00
15. "Computational speech processing," by Conrad F. Sabourin; 2 volumes, 1187 pp. ISBN 2-921173-21-2 and 2-921173-22-0, $150.00
16. "Computational linguistics in information science," by Conrad F. Sabourin; 2 volumes, 1047 pp. ISBN 2-921173-23-9 and 2-921173-24-7, $150.00
17. "Optical character recognition and document segmentation," by Conrad F. Sabourin; 512 pp. ISBN 2-921173-25-5, $80.00