Theory and Applications of Natural Language Processing

Aims and Scope

The field of Natural Language Processing (NLP) has expanded explosively over the past decade: growing bodies of available data, novel fields of applications, emerging areas and new connections to neighboring fields have all led to increasing output and to diversification of research.

"Theory and Applications of Natural Language Processing" is a series of volumes dedicated to selected topics in NLP and Language Technology. It focuses on the most recent advances in all areas of the computational modeling and processing of speech and text across languages and domains. Due to the rapid pace of development, the diversity of approaches and application scenarios are scattered in an ever-growing mass of conference proceedings, making entry into the field difficult for both students and potential users. Volumes in the series facilitate this first step and can be used as a teaching aid, advanced-level information resource or a point of reference.

The series encourages the submission of research monographs, contributed volumes and surveys, lecture notes and textbooks covering research frontiers on all relevant topics, offering a platform for the rapid publication of cutting-edge research as well as for comprehensive monographs that cover the full range of research on specific problem areas.

The topics include applications of NLP techniques to gain insights into the use and functioning of language, as well as the use of language technology in applications that enable communication, knowledge management and discovery such as natural language generation, information retrieval, question-answering, machine translation, localization and related fields.

The books are available in printed and electronic (e-book) form:

* Downloadable on your PC, e-reader or iPad
* Enhanced by Electronic Supplementary Material, such as algorithms, demonstrations, software, images and videos
* Available online within an extensive network of academic and corporate R&D libraries worldwide
* Never out of print thanks to innovative print-on-demand services
* Competitively priced print editions for eBook customers thanks to MyCopy service http://www.springer.com/librarians/e-content/mycopy

For further volumes:
http://www.springer.com/series/8899

David Chiang

Grammars for Language and Genes

Theoretical and Empirical Investigations

Foreword by Aravind K. Joshi

Springer

David Chiang
USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
USA
chiang@isi.edu

Foreword by
Aravind K. Joshi
Department of Computer and Information Science
and
Institute for Research in Cognitive Science
University of Pennsylvania
Philadelphia, PA 19104
USA
joshi@seas.upenn.edu

ISSN 2192-032X e-ISSN 2192-0338
ISBN 978-3-642-20443-2 e-ISBN 978-3-642-20444-9
DOI 10.1007/978-3-642-20444-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011937575

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

JMJ

Foreword

It is indeed a great pleasure to write a few comments on this fascinating and inspiring book: **Grammars for Language and Genes: Theoretical and Empirical Investigations,** by David Chiang. First, I would like to acknowledge my good fortune in being able to work with David during his stay at the University of Pennsylvania. Each one of our meetings was a joyful event, informing and learning from each other, to our mutual benefit.

Chiang's work began with the study of the strong generative capacity of grammars, i.e., their capacity to represent structural descriptions. It is this aspect that is truly important for the study of formal grammars from the perspective of linguistics as well as computational linguistics. However, surprisingly, there is not much work done on issues concerning strong generative capacity (SGC). This is because it is not easy to formulate concepts of SGC that are formal enough and also linguistically meaningful. Building on notions of local interpretation functions, Chiang has given insightful accounts of how SGC should be characterized. He has then applied these ideas to a detailed study of characterizing SGC for a variety of formalisms including tree-adjoining grammars, their variants, and also several other formalisms. Further, building on some notions of extracting more SGC without increasing the weak generative capacity, Chiang has obtained some essential results connecting representations and interpretations. I am confident that much of this work will, in time, become the foundation on which to build further work on the formal characterizations of structural descriptions and interpretations and their eventual use in natural language processing (NLP).

The notion of squeezing more SGC without increasing the weak generative capacity plays a very significant role in the work described in the chapters on statistical parsing and machine translation. These investigations have been carried out in the general framework of tree-adjoining grammar (TAG) and some of its variants. I am sure researchers at large in statistical parsing and machine translation will be inspired by this work and will explore its implications for other classes of formal grammars, thus providing some unity in the very extensive work going on in these areas.

By a remarkable coincidence, just as Chiang was engaged in the activities described above, he also became a member of the group which began to explore the role of formal grammars in characterizing biomolecular structures, such as DNA/RNA and proteins, for example. This part of Chiang's book is a delightful treat for those who want to get a quick but thorough introduction to biomolecular structures and how to model a variety of these structures, keeping both the formal and computational aspects in mind at all times.

In summary, Chiang's work on grammars, which is based on solid mathematical foundations combined with a clear understanding of the domains that are being modeled, will lead to both a deeper theoretical understanding as well as usable computational models. I strongly recommend this book to all those who have already embarked on such activities but, more importantly, to those who would like to be involved in these exciting directions of research.

Aravind K. Joshi

Acknowledgements

This book was originally my PhD dissertation at the University of Pennsylvania. I would like to thank my adviser Aravind Joshi for his guidance during and after graduate school, and Ken Dill at UCSF who was an unofficial co-advisor for Chapters 5 and 6. My dissertation committee, Fernando Pereira, Jean Gallier, Sampath Kannan, and Giorgio Satta, provided a range of insightful perspectives; Giorgio Satta's careful reading was particularly helpful. Other faculty at Penn also gave me much valuable advice: Mark Steedman, Mitch Marcus, Martha Palmer, and David Searls.

Among the graduate students and postdocs I interacted with at Penn, I am particularly indebted to Dan Bikel, Mark Dras, and William Schuler. My collaborations with them and countless discussions taught me half of what I know about computational linguistics. The other half no doubt comes from the rest, including Mike Collins, Hoa Trang Dang, Dan Gildea, Julia Hockenmaier, Seth Kulick, Alexandra Kinyon, Tom Morton, Anoop Sarkar, and Fei Xia.

The revision of the thesis for publication incorporated material from a tutorial on synchronous grammars given with Kevin Knight, which benefited from feedback from Liang Huang and Philip Resnik. During the revision, I received many important comments and corrections from Steve DeNeefe, Jonathan May, and the anonymous reviewer.

I thank my sister Lillian and my future wife Juliana for their loving support, and my parents, who have given me more than I could possibly thank them for.

My graduate research was supported in part by the following grants: ARO DAAG5591-0228, NSF SBR-89-20230, and NSF ITR EIA-02-05456, awarded to the University of Pennsylvania.

Contents