Preface

THE NEED for a thorough textbook for Statistical Natural Language Processing hardly needs to be argued for in the age of on-line information, electronic communication and the World Wide Web. Increasingly, businesses, government agencies and individuals are confronted with large amounts of text that are critical for working and living, but not well enough understood to get the enormous value out of them that they potentially hide.

At the same time, the availability of large text corpora has changed the scientific approach to language in linguistics and cognitive science. Phenomena that were not detectable or seemed uninteresting in studying toy domains and individual sentences have moved into the center field of what is considered important to explain. Whereas as recently as the early 1990s quantitative methods were seen as so inadequate for linguistics that an important textbook for mathematical linguistics did not cover them in any way, they are now increasingly seen as crucial for linguistic theory.

In this book we have tried to achieve a balance between theory and practice, and between intuition and rigor. We attempt to ground approaches in theoretical ideas, both mathematical and linguistic, but simultaneously we try to not let the material get too dry, and try to show how theoretical ideas have been used to solve practical problems. To do this, we first present key concepts in probability theory, statistics, information theory, and linguistics in order to give students the foundations to understand the field and contribute to it. Then we describe the problems that are addressed in Statistical Natural Language Processing (NLP), like tagging and disambiguation, and a selection of important work so

that students are grounded in the advances that have been made and, having understood the special problems that language poses, can move the field forward.

When we designed the basic structure of the book, we had to make a number of decisions about what to include and how to organize the material. A key criterion was to keep the book to a manageable size. (We didn't entirely succeed!) Thus the book is not a complete introduction to probability theory, information theory, statistics, and the many other areas of mathematics that are used in Statistical NLP. We have tried to cover those topics that seem most important in the field, but there will be many occasions when those teaching from the book will need to use supplementary materials for a more in-depth coverage of mathematical foundations that are of particular interest.

We also decided against attempting to present Statistical NLP as homogeneous in terms of the mathematical tools and theories that are used. It is true that a unified underlying mathematical theory would be desirable, but such a theory simply does not exist at this point. This has led to an eclectic mix in some places, but we believe that it is too early to mandate that a particular approach to NLP is right and should be given preference to others.

A perhaps surprising decision is that we do not cover speech recognition. Speech recognition began as a separate field to NLP, mainly growing out of electrical engineering departments, with separate conferences and journals, and many of its own concerns. However, in recent years there has been increasing convergence and overlap. It was research into speech recognition that inspired the revival of statistical methods within NLP, and many of the techniques that we present were developed first for speech and then spread over into NLP. In particular, work on language models within speech recognition greatly overlaps with the discussion of language models in this book. Moreover, one can argue that speech recognition is the area of language processing that currently is the most successful and the one that is most widely used in applications. Nevertheless, there are a number of practical reasons for excluding the area from this book: there are already several good textbooks for speech, it is not an area in which we have worked or are terribly expert, and this book seemed quite long enough without including speech as well. Additionally, while there is overlap, there is also considerable separation: a speech recognition textbook requires thorough coverage of issues in signal analysis and

acoustic modeling which would not generally be of interest or accessible to someone from a computer science or NLP background, while in the reverse direction, most people studying speech would be uninterested in many of the NLP topics on which we focus.

Other related areas that have a somewhat fuzzy boundary with Statistical NLP are machine learning, text categorization, information retrieval, and cognitive science. For all of these areas, one can find examples of work that is not covered and which would fit very well into the book. It was simply a matter of space that we did not include important concepts, methods and problems like minimum description length, backpropagation, the Rocchio algorithm, and the psychological and cognitive-science literature on frequency effects on language processing.

The decisions that were most difficult for us to make are those that concern the boundary between statistical and non-statistical NLP. We believe that, when we started the book, there was a clear dividing line between the two, but this line has become much more fuzzy recently. An increasing number of non-statistical researchers use corpus evidence and incorporate quantitative methods. And it is now generally accepted in Statistical NLP that one needs to start with all the scientific knowledge that is available about a phenomenon when building a probabilistic or other model, rather than closing one's eyes and taking a clean-slate approach.

Many NLP researchers will therefore question the wisdom of writing a separate textbook for the statistical side. And the last thing we would want to do with this textbook is to promote the unfortunate view in some quarters that linguistic theory and symbolic computational work are not relevant to Statistical NLP. However, we believe that there is so much quite complex foundational material to cover that one simply cannot write a textbook of a manageable size that is a satisfactory and comprehensive introduction to all of NLP. Again, other good texts already exist, and we recommend using supplementary material if a more balanced coverage of statistical and non-statistical methods is desired.

STATISTICAL NATURAL LANGUAGE PROCESSING

A final remark is in order on the title we have chosen for this book. Calling the field *Statistical Natural Language Processing* might seem questionable to someone who takes their definition of a statistical method from a standard introduction to statistics. Statistical NLP as we define it comprises all quantitative approaches to automated language processing, including probabilistic modeling, information theory, and linear algebra.

While probability theory is the foundation for formal statistical reasoning, we take the basic meaning of the term 'statistics' as being broader, encompassing all quantitative approaches to data (a definition which one can quickly confirm in almost any dictionary). Although there is thus some potential for ambiguity, Statistical NLP has been the most widely used term to refer to non-symbolic and non-logical work on NLP over the past decade, and we have decided to keep with this term.

Acknowledgments. Over the course of the three years that we were working on this book, a number of colleagues and friends have made comments and suggestions on earlier drafts. We would like to express our gratitude to all of them, in particular, Einat Amitay, Chris Brew, Thorsten Brants, Andreas Eisele, Michael Ernst, Oren Etzioni, Marc Friedman, Éric Gaussier, Eli Hagen, Marti Hearst, Nitin Indurkhya, Michael Inman, Mark Johnson, Rosie Jones, Tom Kalt, Andy Kehler, Julian Kupiec, Michael Littman, Arman Maghbouleh, Amir Najmi, Kris Popat, Fred Popowich, Geoffrey Sampson, Hadar Shemtov, Scott Stoness, David Yarowsky, and Jakub Zavrel. We are particularly indebted to Bob Carpenter, Eugene Charniak, Raymond Mooney, and an anonymous reviewer for MIT Press, who suggested a large number of improvements, both in content and exposition, that we feel have greatly increased the overall quality and usability of the book. We hope that they will sense our gratitude when they notice ideas which we have taken from their comments without proper acknowledgement.

We would like to also thank: Francine Chen, Kris Halvorsen, and Xerox PARC for supporting the second author while writing this book, Jane Manning for her love and support of the first author, Robert Dale and Dikran Karagueuzian for advice on book design, and Amy Brand for her regular help and assistance as our editor.

Feedback. While we have tried hard to make the contents of this book understandable, comprehensive, and correct, there are doubtless many places where we could have done better. We welcome feedback to the authors via email to cmanning@acm.org or hinrich@hotmail.com.

In closing, we can only hope that the availability of a book which collects many of the methods used within Statistical NLP and presents them

in an accessible fashion will create excitement in potential students, and help ensure continued rapid progress in the field.

Christopher Manning

Hinrich Schütze

February 1999