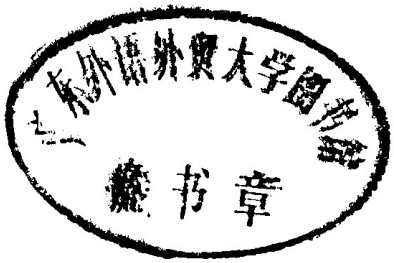
Foundations of Statistical Natural Language Processing

Christopher D. Manning
Hinrich Schütze



The MIT Press
Cambridge, Massachusetts
London, England

Second printing, 1999

Second printing with corrections, 2000

Typeset in 10/13 Lucida Bright by the authors using LaTeX 2ε.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Information

Manning, Christopher D.
Foundations of statistical natural language processing / Christopher D. Manning, Hinrich Schutze.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 0-262-13360-l
1. Computational linguistics-Statistical methods. I. Schutze, Hinrich. II. Title.
P98.5.S83M36 1999
410'.285—dc21 99-21137
CIP

Brief Contents

Contents

List of Tables

List of Figures

Table of Notations

$\cup$	Union of sets
$\cap$	Intersection of sets
A-B, $A \setminus B$	Set difference
$\overline{A}$	The complement of set A
0	The empty set
2^A, $\mathcal{P}(A)$	The power set of A
$\|A\|$	Cardinality of a set
$\sum$	Sum
$\prod$	Product
$p \Rightarrow q$	p implies q (logical inference)
$p \Leftrightarrow q$	p and q are logically equivalent
$\stackrel{\text{def}}{=}$	Defined to be equal to (only used if "=" is ambiguous)
$\mathbb{R}$	The set of real numbers
$\mathbb{N}$	The set of natural numbers
$n!$	The factorial of n
∞	Infinity
$\|x\|$	Absolute value of a number
$\ll$	Much smaller than
$\gg$	Much greater than
$f : A \to B$	A function f from values in A to B
$\max f$	The maximum value of f

$\min f$	The minimum value of f
$\arg\max f$	The argument for which f has its maximum value
$\arg\min f$	The argument for which f has its minimum value
$\lim_{x\to\infty} f(x)$	The limit of f as x tends to infinity
$f \propto g$	f is proportional to g
∂	Partial derivative
$\int$	Integral
$\log a$	The logarithm of a
$\exp(x), e^x$	The exponential function
$\lceil a \rceil$	The smallest integer i s.t. $i \geq a$
$\vec{x}$	A real-valued vector: $\vec{x} \in \mathbb{R}^n$
$\lvert\vec{x}\rvert$	Euclidean length of $\vec{x}$
$\vec{x} \cdot \vec{y}$	The dot product of $\vec{x}$ and $\vec{y}$
$\cos(\vec{x}, \vec{y})$	The cosine of the angle between $\vec{x}$ and $\vec{y}$
c_{ij}	Element in row i and column j of matrix C
C^{T}	Transpose of matrix C
$\hat{X}$	Estimate of X
$E(X)$	Expectation of X
$\mathrm{Var}(X)$	Variance of X
μ	Mean
σ	Standard deviation
$\bar{x}$	Sample mean
s^2	Sample variance
$P(A\|B)$	The probability of A conditional on B
$X \sim p(x)$	Random variable X is distributed according to p
$b(r; n, p)$	The binomial distribution
$\binom{n}{r}$	Combination or binomial coefficient (the number of ways of choosing r objects from n)
$n(x; \mu, \sigma)$	The normal distribution
$H(X)$	Entropy

$I(X; Y)$	Mutual information
$D(\mathrm{p} \| \mathrm{q})$	Kullback-Leibler (KL) divergence
$C(\cdot)$	Count of the entity in parentheses
f_u	The relative frequency of u.
w_{ij}, W(i)(j)	The words $w_i, w_{i+1}, \ldots, w_j$
$w_{i,j}$	The same as w_{ij}
$w_i, \ldots, w_j$	The same as w_{ij}
$O(n)$	Time complexity of an algorithm
*	Ungrammatical sentence or phrase or ill-formed word
?	Marginally grammatical sentence or marginally acceptable phrase

Note. Some chapters have separate notation tables for symbols that are used locally: table 6.2 (Statistical Inference), table 7.1 (Word Sense Disambiguation), table 9.1 (Markov Models), table 10.2 (Tagging), table 11.1 (Probabilistic Context-Free Grammars), and table 14.2 (Clustering).

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, back-propagation, 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, Eric 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 Schiitze
February 1999

Road **Map**

IN GENERAL, this book is written to be suitable for a graduate-level semester-long course focusing on Statistical NLP. There is actually rather more material than one could hope to cover in a semester, but that richness gives ample room for the teacher to pick and choose. It is assumed that the student has prior programming experience, and has some familiarity with formal languages and symbolic parsing methods. It is also assumed that the student has a basic grounding in such mathematical concepts as set theory, logarithms, vectors and matrices, summations, and integration - we hope nothing more than an adequate high school education! The student may have already taken a course on symbolic NLP methods, but a lot of background is not assumed. In the directions of probability and statistics, and linguistics, we try to briefly summarize all the necessary background, since in our experience many people wanting to learn about Statistical NLP methods have no prior knowledge in these areas (perhaps this will change over time!). Nevertheless, study of supplementary material in these areas is probably necessary for a student to have an adequate foundation from which to build, and can only be of value to the prospective researcher.

What is the best way to read this book and teach from it? The book is organized into four parts: Preliminaries (part I), Words (part II), Grammar (part III), and Applications and Techniques (part IV).

Part I lays out the mathematical and linguistic foundation that the other parts build on. Concepts and techniques introduced here are referred to throughout the book.

Part II covers word-centered work in Statistical NLP. There is a natural progression from simple to complex linguistic phenomena in its four

chapters on collocations, n-gram models, word sense disambiguation, and lexical acquisition, but each chapter can also be read on its own.

The four chapters in part III, Markov Models, tagging, probabilistic context free grammars, and probabilistic parsing, build on each other, and so they are best presented in sequence. However, the tagging chapter can be read separately with occasional references to the Markov Model chapter.

The topics of part IV are four applications and techniques: statistical alignment and machine translation, clustering, information retrieval, and text categorization. Again, these chapters can be treated separately according to interests and time available, with the few dependencies between them marked appropriately.

Although we have organized the book with a lot of background and foundational material in part I, we would not advise going through all of it carefully at the beginning of a course based on this book. What the authors have generally done is to review the really essential bits of part I in about the first 6 hours of a course. This comprises very basic probability (through section 2.1.8), information theory (through section 2.2.7), and essential practical knowledge – some of which is contained in chapter 4, and some of which is the particulars of what is available at one's own institution. We have generally left the contents of chapter 3 as a reading assignment for those without much background in linguistics. Some knowledge of linguistic concepts is needed in many chapters, but is particularly relevant to chapter 12, and the instructor may wish to review some syntactic concepts at this point. Other material from the early chapters is then introduced on a "need to know" basis during the course.

The choice of topics in part II was partly driven by a desire to be able to present accessible and interesting topics early in a course, in particular, ones which are also a good basis for student programming projects. We have found collocations (chapter 5), word sense disambiguation (chapter 7), and attachment ambiguities (section 8.3) particularly successful in this regard. Early introduction of attachment ambiguities is also effective in showing that there is a role for linguistic concepts and structures in Statistical NLP. Much of the material in chapter 6 is rather detailed reference material. People interested in applications like speech or optical character recognition may wish to cover all of it, but if n-gram language models are not a particular focus of interest, one may only want to read through section 6.2.3. This is enough to understand the concept of likelihood, maximum likelihood estimates, a couple of simple smoothing methods (usually necessary if students are to be building any

probabilistic models on their own), and good methods for assessing the performance of systems.

In general, we have attempted to provide ample cross-references so that, if desired, an instructor can present most chapters independently with incorporation of prior material where appropriate. In particular, this is the case for the chapters on collocations, lexical acquisition, tagging, and information retrieval.

Exercises. There are exercises scattered through or at the end of every chapter. They vary enormously in difficulty and scope. We have tried to provide an elementary classification as follows:

* Simple problems that range from text comprehension through to such things as mathematical manipulations, simple proofs, and thinking of examples of something.

** More substantial problems, many of which involve either programming or corpus investigations. Many would be suitable as an assignment to be done over two weeks.

★★★ Large, difficult, or open-ended problems. Many would be suitable as a term project.

Website. Finally, we encourage students and teachers to take advantage of the material and the references on the companion *website*. It can be accessed directly at the URL http://www.sultry.arts.usyd.edu.au/fsnlp, or found through the MIT Press website http://mitpress.mit.edu, by searching for this book.

WEBSITE