まえがき

オンライン情報，電気通信，そして World Wide Web の時代において，統計的自然言語処理を徹底的に論じた教科書が必要であることには議論の余地がない．企業も政府機関も個人も，ますます大量の，業務や生活に欠かせないテキスト情報に直面しているのに，それらが潜在的に隠し持っている莫大な価値を引き出す術を十分理解できずにいる．

同時に，巨大なテキストコーパスが入手できるようになったことで，言語学や認知科学における言語への科学的方法論も変化している．人工的で小さな領域や個々の文を研究していたのでは検知できなかった，あるいは興味深いと思われなかった現象が，説明されるべき重要な事項の中核に来つつある．ほんのちょっと以前，1990 年代においてさえ，定量的な手法は言語学において不適切とされ，数理言語学の主たる教科書でさえそれをまったく扱わなかったのであるが，現在，言語学理論において，それら定量的な手法は徐々に不可欠なものと認識されつつある．

本書の執筆において，筆者らは，理論と実践，直観と厳密さの釣り合いをとることに尽力してきた．さまざまな手法を，数学のそして言語学の理論的な発想に基礎付けようとしてきたが，それと同時に内容があまりにも無味乾燥にならないようにし，理論的な発想が実用的な問題を解決するにどのように役立つのかを示すようにした．これを実現するために，まず，この分野を理解するのに必要であり，そこに寄与するための基礎を読者である学生たちに与えることとし，確率論，統計学，情報理論，言語学の主要な概念を先に示している．その後，タグ付けや曖昧性解消など統計的自然言語処理が取り組んでいる問題と，重要な研究を選んで，記述している．これによって，学生たちは，すでに示されている基礎に根ざすことができ，言語がもたらす特別の問題も理解しているので，この分野を前に進んでいけるようになると考えている．

本書の基本的な構成を設計した際，何を含め，それらの素材をどう整理する

かについて，多くの決定をしなければならなかった．主たる基準は，本の厚さを扱える程度に納めることであった（完璧に成功したとは言い難い！）．そのため，本書では，確率論，情報理論，統計学，そして統計的自然言語処理で用いられるその他の多くの数学の分野について，完全な紹介は行えていない．この分野で最も重要と思われる話題は取り上げるようにしたが，本書を用いて授業を行う人たちは，特に興味のある数学的基礎をより深く理解させるために，補助的な教材を使う必要がある場合も多々あると思う．

また，統計的自然言語処理を，そこで用いられる数学的道具立てや理論に関して，一つの統一的な枠組みだけを用いて提示しようという試みもしないこととした．統一的な数学の理論的基盤が望ましいことは間違いないが，現時点では，そのような理論はそもそも存在しない．このため，いくつかの場所で多くの要素が折衷的に入り交じることになっているが，自然言語処理に対して特定の方法論が正しく，それ以外のものよりも優先すると決めつけるのは，まだ時期尚早と考えている．

ことによると驚かれるかもしれないが，音声認識は扱わないこととした．音声認識は，自然言語処理とは異なる分野として始められ，主に電気工学科で育ち，異なる会議と学会誌を持ち，それ独自の関心を多く抱えている．とはいえ，最近では，ますますの融合と重なり合いが進んでいる．自然言語処理において統計的方法論の復活を引き起こしたのは，音声認識に関する研究であったし，本書で示す手法の多くは最初音声のために開発され，自然言語処理の中に広まっていった．特に，音声認識における言語モデルの研究は本書における言語モデルの議論と広い重なりを持つ．しかも，音声認識は，現在最も成功していて，最も広く応用されている言語処理の分野であると主張することもできる．しかし，この分野を本書に含めなかった実際的な理由がいくつかある．音声については優れた教科書が数多く存在する，この分野は著者らが研究を続け，本当の専門家となったものではない，そして，本書は音声を含めなくてもすでに十分に長いのである．さらに，重なりもあるとはいえ，自然言語処理と音声認識の間には重要な違いもある．音声認識の教科書は信号解析と音響モデルについてひととおり取り上げる必要がある．これらは普通，計算機科学や自然言語処理を背景としている人たちにとって興味がわかず，身近でもない内容である．逆に，音声を学んでいる人たちほとんどは，本書で注目している多くの自然言語処理の話題には興味がないだろう．

統計的自然言語処理との境界がある意味不明瞭で，それと関連を持つそれ以外の領域として，機械学習，テキスト分類，情報検索，そして認知科学があげられる．これらの領域すべてにおいて，本書では取り扱ってはいないが，内容的には本書に適当であるような研究の例を見つけることができる．最小記述長，誤差逆伝搬法，Rocchio アルゴリズム，そして，言語処理における出現頻度効

果についての心理学的，認知科学的文献等々，多くの重要な概念，手法，問題が本書に含まれていないのは，単に分量からの理由である．

著者らにとって下すのが最も難しかった決定は，統計的自然言語処理と統計的ではないそれとの境界に関するものであった．著者らが本書の執筆を始めた頃には，この二つには明確な境界線があったが，最近それがどんどんはっきりしないものになってきているように思う．統計的でない研究者もコーパスを利用し，定量的な手法を取り入れることが多くなっている．そして，統計的自然言語処理において，確率的モデルやその他のモデルの構築を始めるにあたっては，目をつぶっての白紙の状態からではなく，現象について明らかになっている科学的知識はすべて利用する必要があることが，今では一般的に認められている．

そういうわけであるので，多くの自然言語処理研究者が，統計的側面だけを取り上げた教科書を書くことが良識的かと疑問を感じることと思う．そして，著者らが最も望まないことは，この教科書が，一部の人たちが持っている，統計的自然言語処理には言語理論や記号計算の研究は関係ないのだという不幸な見方を助長してしまうことである．とはいえ，複雑で本当に多くの基礎的な事柄があり，自然言語処理のすべてへの導入となるような十分かつ網羅的な教科書を，扱える厚さに納めることはまったく不可能なのだと考えている．繰り返しになるが，素晴らしい教科書は既に何冊もある．統計的な手法と統計的でない手法のより釣り合いのとれた扱いが必要な場合は，それらを補助的な教材として用いていただくことをお薦めする．

統計的自然言語処理（STATISTICAL NATURAL LANGUAGE PROCESSING）

最後に本書のために選んだ書名の適切さについて言及しておく．ここで扱っている分野を**統計的自然言語処理** (*statistical natural language processing*) と呼ぶことは，統計学における標準的な入門書から統計的手法の定義を持ってきている人たちにとっては疑問のあるものと思う．ここで著者らが定義する統計的自然言語処理とは，自動言語処理における定量的な手法すべてからなる．そこには，確率的モデル，情報理論，線形代数も含まれる．確率論は形式的な統計的推論の基礎であるが，著者らは「統計」という用語の意味をより広いものと捉え，データに対するすべての定量的手法をそこに含めている（ほとんどの辞書で誰でも簡単に確認できる定義である）．そのため，曖昧さが生じる可能性はあるのだが，統計的自然言語処理は，ここ 10 年以上にわたって，自然言語処理において，記号的でなく，論理に偏らない研究を参照するために広く使われてきている用語であるため，著者らもこの用語を使い続けることとする．

謝辞：本書を執筆していた 3 年間の間，多くの同僚と友人たちから，初期の草稿へコメントや提案を頂いた．彼ら全員に感謝を表したい．特に，Einat Amitay, Chris Brew, Thorsten Brants, Gary Cottrell, 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. 著者らは特に，Bob Carpenter, Eugene Charniak, Raymond Mooney, そして MIT 出版の匿名の査読者に恩義を感じている．彼らは，内容と説明方法の両方について，多くの改善を提案してくださり，そのお陰で，本書の質や使いやすさは大きく向上したように感じている．自分たちのコメントから導かれたアイディアに気づかれたときは，それぞれに謝辞はなくても，著者らが感謝していることを彼らに感じとっていただければと思う．

さらに，本書を執筆中の第二著者を支援してくださったことに対して，Francine Chen, Kris Halvorsen，そして，Xerox PARC に，第一著者への愛と支援に対して，Jane Manning に，本書の装丁に助言いただいた Robert Dale と Dikran Karagueuzian に，編集者としてずっと助けてくれた Amy Brand に，感謝を捧げたい．

ご意見ご批評：本書の内容を，理解しやすく，包括的で，正しいものにしようと，十分努力したつもりだが，もっとうまくできたはずの箇所が数多くあるのは間違いないことと思う．ご意見やご批評などは電子メールにて，cmanning@acm.org か me@hinrichschuetze.com に送ってほしい．

最後に，統計的自然言語処理で用いられている多くの手法を集め，それらをわかりやすく説明している本書の登場によって，現在そして将来の学生たちが多くの刺激を受け，この分野における急速で継続的な進歩を確かなものにすることの助けになること，それだけを希望する．

Christopher Manning

Hinrich Schütze

February 1999

統計的自然言語処理の基礎
Christopher D.Manning ・Hinrich Schutze 著・加藤 恒昭・菊井 玄一郎・林 良彦・森 辰則訳
http://www.kyoritsu-pub.co.jp/bookdetail/9784320124219

本書の使い方

おおよそのところ，本書は，統計的自然言語処理に焦点を当てた大学院レベルの 1 セメスタの講義に適当となるように執筆されている．実際には，1 セメスタで扱おうと思うよりはやや多めの素材が含まれている．ただ，そのような豊富さは，教師に取捨選択の大きな余地を与えると思う．学生については，事前に，プログラミングの経験を持ち，形式言語と記号的な構文解析手法にある程度の馴染みがあることを想定している．さらに，集合論，対数，ベクトルと行列，積和，積分などの数学的概念について初歩的な基礎を身につけていると考えている．ただし，高校を卒業した学生が適切に身につけているだろうもの以上を望むわけではない．学生は記号的自然言語処理の授業を先に受けていてももちろん構わないが，その基礎の多くを前提とするようなことはしていない．確率論，統計学，言語学については，必要な背景の簡単な要約を本書に含めている．著者らの経験では，統計的自然言語処理手法を学ぼうとする人たちの多くがこれらの分野に関しての知識を事前に持っていないことが多いためである（たぶん，時が経てば状況は変わると思うが）．とはいえ，補助的な教材を用いてこれらの分野について学ぶことは，それを構築している適切な基盤を身につけるために，学生にとっておそらく必要だろうし，将来の研究者になるためにも役に立つこととなるだろう．

本書を読むのに，そしてこれを用いて授業をするのに最も良い方法を考えてみる．本書は四つの部分からなっている．前提知識（I 編），語（II 編），文法（III 編），そして応用と技法（IV 編）である．

前提知識（I 編）は，その他の部分の前提となる数学的，言語学的基礎を配置している．ここで導入される概念と技法は，本書全般を通じて参照される．

語（II 編）は，統計的言語処理における語を中心とした研究を扱っている．連語，n–グラムモデル，語義曖昧性解消，語彙獲得の四つの章からなるが，これらは単純なものから複雑な言語現象へと自然に進むように配置されている．た

だし，それぞれの章は独立して読むことが可能である．

文法（III 編）の四つの章は，マルコフモデル，タグ付け，確率文脈自由文法，そして，確率的構文解析であるが，お互いに依存しているので，順序よく教えていった方がよい．ただし，タグ付けの章は，マルコフモデルの章を時々参照する必要があるとはいえ，独立したものとして読むことができる．

応用と技法（IV 編）の題材は，統計的アライメントと機械翻訳，クラスタリング，情報検索，テキスト分類という，四つの応用と技術である．これらの章も，いくつかあるお互いの関係について適当に言及していただきつつ，興味と許された時間に応じて，別々に扱うことができる．

本書では，背景もしくは基礎となる多くの材料を I 編にまとめるという構成をとっているが，本書に基づいて授業を行う際に，その先頭でそれらすべてを丁寧に説明していくことはお薦めしない．著者らが一般にとっている方法では，講義の最初のおよそ 6 時間で I 編の中の本当に中心的な部分を復習するに留めている．その中には，確率論の本当の基礎（2.1.8 節まで），情報理論（2.2.7 節まで），そして基本的な実用的知識が含まれる．この実用的知識の一部は 4 章に含まれているし，その他はそれぞれの組織が何を持っているかに応じた個別事項となろう．著者らは，普通，言語学の背景をあまり持たない学生に対して，3 章の内容は読んでおくようにと宿題として残すようにしている．言語学的な概念に関するいろいろな知識は多くの章で必要となるが，12 章は特にそれらと関連が深く，教師はこの時点で統語に関連する概念を復習したくなるのではと思う．最初の方の節のこれら以外の素材については，授業を通じて，「知るのが必要になったら」という基準で導入していけばよいだろう．

II 編の話題は，馴染みがあって興味深い話題，特に学生のプログラミング課題の良い基礎となるようなものを授業の最初の方で提示できるようにしたいという要望に一部導かれている．連語関係（5 章），語義曖昧性解消（7 章），付加の曖昧性（8.3 節）が，この点で特にうまくいくことに気がついている．付加の曖昧性を早めに扱うことは，統計的自然言語処理において言語学的な概念や構造に役割があることを示すのにも効果的である．6 章の内容の多くは，比較的詳細で参考資料的な素材である．音声認識や光学的文字認識などの応用に興味のある人たちはこれら全部を扱おうと思うだろうが，n–グラムの言語モデルが興味の焦点でないのであれば，6.2.3 節まで読む程度でよいと思うかもしれない．尤度や最尤推定の概念，いくつかのスムージング手法（学生が自分自身の確率モデルを構築する際には普通必要となる），そしてシステムの性能を評価する適切な手法を理解するには，それで十分である．

全般にわたって多くの相互参照を行うように努めたので，もし望むのであれば，ほとんどの章は独立したものとして教えることができる，適当なところで，参照されている以前の素材を含めていくようにすればよい．このことは，特に，連語，語彙獲得，タグ付け，情報検索の章にあてはまる．

練習問題：各章のそこかしこ，または章末に練習問題をおいている．これらはその難しさも扱う範囲も本当にさまざまである．以下のようなおおざっぱな分類を試みている．

⋆ 簡単な問題で，本文の理解を問うものから，数学的な変形操作，簡単な証明，何かの例を考えるような問題までに及ぶ．

⋆⋆ より実質的な問題で，多くはプログラミングかコーパス調査を含んでいる．これらの多くは 2 週間以上の期間で行う課題とするのが適当である．

⋆⋆⋆ 大きく，困難で，広がりを持ち，決まった答えのない問題である．多くは，学期末課題とするのが適当である．

Web サイト (website)

Web サイト：最後に，学生と教師のみなさんに，参考**Web サイト** (*website*) にある素材や参考文献を活用していただくことをお薦めする．http://nlp.stanford.edu/fsnlp の URL で直接アクセスすることもできるし，MIT 出版の Web サイト http://mitpress.mit.edu で，本書を探してそこから辿ることもできる．