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

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.