Linguistic Structure Prediction

Noah A. Smith
Carnegie Mellon University

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #13

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

A major part of natural language processing now depends on the use of text data to build linguistic analyzers. We consider statistical, computational approaches to modeling linguistic structure. We seek to unify across many approaches and many kinds of linguistic structures. Assuming a basic understanding of natural language processing and/or machine learning, we seek to bridge the gap between the two fields. Approaches to decoding (i.e., carrying out linguistic structure prediction) and supervised and unsupervised learning of models that predict discrete structures as outputs are the focus. We also survey natural language processing problems to which these methods are being applied, and we address related topics in probabilistic inference, optimization, and experimental methodology.

KEYWORDS

natural language processing, computational linguistics, machine learning, decoding, supervised learning, unsupervised learning, structured prediction, probabilistic inference, statistical modeling

Contents

Preface

The title of this book is ambiguous. The intended reading involves structure of a linguistic nature (*linguistic structure*) that is to be predicted. This is the meaning implied by the disambiguated representation below:

$$[\,[\,\text{linguistic structure}\,]_{\text{noun phrase}}\ \text{prediction}\,]_{\text{noun phrase}}$$

By the noun phrase *linguistic structure*, we refer to symbolic representations of language posited by some theory of language. The representation above is a linguistic structure. The noun phrase *linguistic structure prediction* refers to automatic methods for annotating, analyzing, or disambiguating text.

The alternative reading of the title attaches the adjective *linguistic* to *prediction*, leaving it unclear what kind of structure is to be predicted, but suggesting that the method of prediction involves linguistics. The ambiguity in the title serves as a reminder that ambiguity is ubiquitous in human language, especially from the point of view of computer programs, and underlies much of the challenge of automatic natural language processing. The title is designed to evoke the three intellectual strands tied together in this volume.

Statistics: The word *prediction* suggests reasoning about something unknown or invisible. The techniques presented here aim to take electronic text data[1] as input and provide as output a hypothesized analysis of the text. Formulating this as a prediction problem leads naturally to the use of statistical models and methods that use past experience, or exposure to data, to make new predictions.

Computer science: The word *structure* evokes ideas about complexity and interconnectedness; in machine learning the term *structure prediction* (or *structured prediction*) is used to refer to prediction of a set of interrelated variables. Such problems arise in areas like computer vision (e.g., interpreting parts of a visual scene), computational biology (e.g., modeling how protein molecules fold), and, of course, speech and text processing. Here, we are interested in discrete structures that can be defined succinctly in mathematical terms and manipulated efficiently by computer programs.

Linguistics: The word *linguistic*, of course, refers to the fact that those discrete structures are stipulated by some theory of human languages. Linguistic structure prediction is perhaps most strongly associated with parsing sentences into syntactic structures as used in theories accounting for well-formedness of some conceivable sentences vis-à-vis others, but the techniques are

[1] Although we assume text to be a sequence of orthographic symbols, the methods here are also relevant when our starting point is even more raw, such as images of text written on a page or recordings of speech.

applicable at other levels of linguistic analysis as well: phonology, morphology, semantics, discourse, and so on. Linguistics as a field of inquiry is notable for the ways in which computational models have influenced theories.

In summary, then, this is a book about machine learning (ML), natural language processing (NLP), and computational linguistics (CL),[2] though it does not claim to cover any of them completely or in a balanced way. We aim for neither the union nor the intersection, but rather a useful and coherent selection of important and useful ideas that are best understood when synthesized.

HISTORICAL CONTEXT

Having parsed the title, we next attempt to explain the existence of this book.

In the past decade, NLP and ML have become increasingly interested in each other. The connection between the broader fields of linguistics and statistics is much older, going back to linguists like George Kingsley Zipf and Zellig Harris who emphasized the centrality of *data* in reasoning about the nature of language, and also to Claude Shannon and the information theorists of the post-World War II era who saw language processing problems like machine translation as problems of decryption that called for statistical reasoning (Weaver, 1949).

The trends of each field have recently deepened this connection substantially. In NLP, the rise in availability of enormous amounts of diverse, multilingual text data, due largely to the rise of the web, has opened the question of just how far data-driven approaches to text processing can go. In ML, advances in graphical models for probabilistic reasoning have permitted the exploration of learning problems with high-dimensional outputs far beyond the classical prediction problems addressed by regression and binary classification. Both fields have been enabled greatly by continued increases in the computational power of computers, of course.

Occasionally tension appears to arise between the field of linguistics and the use of data, particularly among NLP researchers. Linguistics-firsters have seen data-driven methods as "shallow" or, worse, unprincipled and ignorant of linguistic theory. Data-firsters have derided linguistics-based systems as "brittle" and labor-intensive, often questioning the helpfulness of representations that require heavy processing of the raw, "pure" data. Here we reject the linguistics-or-statistics choice as a false one that is based on uncharitably narrow views of what constitutes "data" and what constitutes "linguistics." The preponderance of evidence suggests that NLP works better when it uses data, and that ML works better when it takes into account domain knowledge. In NLP, "domain knowledge" often means linguistics. We take the view that linguistic knowledge is crucial in defining representations of NLP problems and efficient algorithms for handling those representations. Even when it is not acknowledged, it is implicit in NLP researchers' design decisions whenever we build

[2] We will use the term *NLP* throughout. Unfortunately there is no consensus on how this term should be defined. We use it here to refer to the class of computational problems of textual or linguistic analysis, generation, representation, or acquisition. This might include speech processing or text retrieval, though we do not address those here. Some researchers in the field distinguish between NLP and CL, using the former to emphasize engineering goals and the latter to emphasize the *human* language faculty and the use of computational models to understand the nature of language as a natural phenomenon. We do not take a position on whether NLP and CL are the same.

software that handles text. We do not claim that linguistic processing by machines should resemble linguistic processing by humans, but rather that scientific theories about linguistic data—the by-product of human linguistic behavior—offer useful insights for automatic linguistic processing. Of course, linguists do not generally keep NLP in mind as they do their work, so not all of linguistics should be expected to be useful.

Meanwhile, statistics and machine learning offer elegant, declarative formalisms and effective algorithms for transforming raw text data into forms that can support useful, text-related applications. While the abstractions of ML are attractive and show promise for general solutions that can be used in many fields, these abstractions cannot be applied to language problems in ignorance, and so in this book we keep NLP in mind as we explore these abstractions.[3]

WHAT TO EXPECT

This volume aims to bridge the gap between NLP and ML. We begin in chapter 1 by arguing that many central problems in NLP can be viewed as problems of structure prediction (i.e., reasoning about many interdependent events). We hope to demystify some of the core assumptions of NLP for ML researchers, explaining certain conventions and tradeoffs. In chapter 2, we turn to decoding: algorithmic frameworks for making predictions about the linguistic structure of an input. Decoding algorithms often shape the way NLP researchers think about their problems, and understanding the connections among algorithms may lead to more freedom in designing solutions. In chapters 3 and 4, we aim to provide a survey of current ML techniques suitable for linguistic structures. The goal is a unified treatment that reveals the similarities and connections among the many modern approaches to statistical NLP. Despite the complex history leading up to this set of techniques, they are, in many cases, quite similar to each other, though the differences are important to understand if one is to meaningfully use and extend them. Chapter 5 considers important inference problems that arise in NLP, beyond decoding, and the appendices consider topics of practical or historical interest.

The book assumes a basic understanding of probability and statistics, and of algorithms and data structures. The reader will be better prepared if she has taken an introductory course in ML and/or NLP, or read an introductory textbook for either (e.g., Bishop, 2006, for ML, Jurafsky and Martin, 2008, for NLP, and Manning and Schütze, 1999 for statistical NLP). The book was written with three kinds of readers in mind: graduate students, NLP researchers wanting to better understand ML, and ML researchers wanting to better understand NLP. Basic familiarity with NLP and ML is assumed, though probably only one or the other suffices.

WHAT NOT TO EXPECT

Linguists will find nothing here about specific linguistic theories, and perhaps too much about shallow and syntactic models of language. This is simply because so much remains to be done in the

[3] Given the accessibility of NLP problems, we suspect the book may be useful for researchers seeking to use ML for structured problems in other domains, as well. We all do know at least a little about language.

development of learnable, "deep" linguistic models. We hope this book will shed some light on how that development can take place and inspire more linguists to consider how interesting linguistic effects can be described in modern computational frameworks (e.g., feature-based statistical models) and then implemented through the use of appropriate algorithms. Further, we invite the challenge to find cases where the current framework is truly insufficient to model linguistic phenomena of interest, and note that methods that use "incomplete data" may provide a principled starting point for anyone who is unsatisfied with what is visible in data alone.

ML researchers may already be familiar with many of the modeling techniques in chapters 3 and 4, and may wonder why these models' application to language is any different. The short answer is to consider chapter 2 and note the challenges of doing inference efficiently in the presence of features that try to capture linguistic phenomena. Neither the imagination of linguists nor the nature of text data can be confined to the classical formalisms of machine learning.

Engineers may be annoyed at the level of abstraction; very few algorithms are presented for direct implementation. We often deliberately use declarative problem statements rather than procedural algorithms. This is partly for pedagogical reasons, and partly because this book is written primarily for researchers, not developers. In a similar vein, researchers in machine translation, an application of NLP that has become increasingly important and increasingly engineering-focused, may question the relevance of this book for their problems. To date, the cross-fertilization between translation research and other NLP research has been tremendous. Examples from translation are occasionally mentioned, but we do not believe the problem of translation should be treated in a fundamentally different way from other linguistic structure prediction problems. Source-language input and target-language output sentences are both linguistic structures. We hope that the material here will be useful to translation researchers seeking to better exploit ML.

A major topic that this book does not address directly is the scalability of learning algorithms. At this writing, this is an area of intense activity. Our view is that "scaling up" should not require a fundamental change in representations or the choice of learning algorithms, but rather innovation of new methods of approximation. Hence approximations that permit us to efficiently exploit very large datasets or large computer clusters should be considered orthogonal to the kinds of models we seek to learn, with the possible benefit of allowing more powerful models to be built. Often, large amounts of data are found to improve the performance of simpler models and learning approaches, relative to more complex ones that appear superior in smaller-data settings. Since very large amounts of data are not available for all languages, domains, and tasks, well-founded ML and structured NLP are still worthwhile research investments. In short, it is quite possible that scalability concerns will lead to a paradigm change in ML and NLP, but at this writing, we are unsure.

Computational linguists who work on natural language generation may be frustrated by the emphasis on analysis problems (e.g., parsing and sequence models). At a high level of abstraction, analysis and generation are inverses: analysis predicts linguistic structure from surface text (itself a sequential structure), and generation predicts surface text from linguistic structure (or, in some cases, other text). It is perhaps desirable to characterize natural language structure in such a way that the solutions to analysis and generation can be described as inverse operations (e.g., as with finite-state transducers). It remains to be seen whether linguistic structure prediction is a viable approach to generation.

ROAD MAP

The material here roughly corresponds to an advanced graduate course taught at Carnegie Mellon University in 2006–2009, though largely reordered. There are five main chapters and four appendices:

- Chapter 1 is a survey of many examples of linguistic structures that have been considered as outputs to be predicted from text.
- Chapter 2 presents the linear modeling framework and discusses a wide range of algorithmic techniques for predicting structures with linear models, known as "decoding."
- Chapter 3 discusses learning to predict structures from training data consisting of input-output pairs. This is known as supervised learning.
- Chapter 4 turns to learning when the training data are inputs only (no outputs), or otherwise incomplete. We consider unsupervised and hidden variable learning. (Semisupervised learning is not covered.)
- Chapter 5 discusses statistical reasoning algorithms other than decoding, many of which are used as subroutines in the learning techniques of chapters 3 and 4.

The first two appendices discuss some loose ends of practical interest:

- Numerical optimization algorithms (A) and
- Experimentation methods (B).

The remaining appendices are of mainly historical interest:

- A discussion of "maximum entropy" and its connection to linear models (C) and
- A discussion of locally normalized conditional models (D).

Acknowledgments

The development of this book was supported in part by the National Science Foundation through grants IIS-0915187 and CAREER IIS-1054319.

This book is the outcome of my interactions with many people. I thank first those who have taken the "Language and Statistics II" course (11-762) at Carnegie Mellon University in 2006–2009. This monograph reflects much of the dialogue in those lectures. I thank ARK members and alumni: Desai Chen, Shay Cohen, Dipanjan Das, Chris Dyer, Philip Gianfortoni, Kevin Gimpel, Michael Heilman, André Martins, Zack McCord, Daniel Mills, Behrang Mohit, Brendan O'Connor, Naomi Saphra, Nathan Schneider, Cari Sisson, Dan Tasse, Matthew Thompson, Mengqiu Wang, Tae Yano, and Dani Yogatama. I have learned at least as much from you about how to tell this story as you did about the story, and I hope this artifact will serve as a useful reference in your future work.

The project that became this monograph started as a tutorial at the 26th International Conference on Machine Learning in Montréal, June 2009. Thanks to Jen Neville for suggesting the tutorial and the audience for questions and conversations. Thanks to Graeme Hirst for convincing me to expand the tutorial into this volume and for his patience and advice during the writing. Thanks to Mike Morgan and C.L. Tondo for their encouragement, patience, and gentle introduction to the world of book publishing.

The way of thinking presented in this monograph is my own, but it is largely owed to the intellectual influences of Fred Jelinek (a giant now sorely missed) and to having worked closely with Jason Eisner and David Smith. They no doubt would have done it differently, but I hope they'd rather have this monograph than none at all.

This monograph was also shaped by (often subtly, sometimes radically) alternative ways of thinking, and for exposure to many of those over the years, I thank Pedro Aguiar, Jason Baldridge, Alan Black, David Blei, Jamie Callan, Jaime Carbonell, Eugene Charniak (thanks especially for Charniak, 1993), William Cohen, Michael Collins, Hal Daumé, Jacob Eisenstein, Mário Figueiredo, Sharon Goldwater, Carlos Guestrin, Rebecca Hwa, Mark Johnson, Sanjeev Khudanpur, Gary King, Dan Klein, John Lafferty, Alon Lavie, Lori Levin, Chris Manning, Dan Melamed, Tom Mitchell, Ray Mooney, Kemal Oflazer, Miles Osborne, Fernando Pereira, Philip Resnik, Roni Rosenfeld, Dan Roth, Bryan Routledge, Mark Steedman, Stephan Vogel, Bonnie Webber, Jan Wiebe, Eric Xing, and David Yarowsky. I thank Phil Blunsom, Shay Cohen, Chris Dyer, Jacob Eisenstein, Jenny Finkel, Kevin Gimpel, Michael Heilman, Adam Lopez, André Martins, Brendan O'Connor, Nathan Schneider, David Smith, and the anonymous reviewers for comments on drafts. All errors are, of course, solely the responsibility of the author.

And thanks to Karen Thickman for her patience and support; she only reminded me once that I swore, after finishing my thesis, that I'd never do anything like that again.

This volume is dedicated (i) to the memory of my father, Wayne Smith (1943–2010), to whom I owe more than can be expressed in words, and who would have loved to have seen this project completed, (ii) in honor of Julie Brahmer, the physician whose work extended by a precious 20% the time I was able to share with him, and (iii) in honor of my sister, Maria Smith, who is wise and gentle beyond her years and inspires me to believe that Dad is still with us.

Noah A. Smith
Pittsburgh
May 2011