Book Review

Natural Language Processing: A Machine Learning Perspective

Yue Zhang and Zhiyang Teng
(Westlake University)

Cambridge University Press, 2021, ix+470 pp; hardbound, ISBN 978-1-108-42021-1;
doi:10.1017/9781108332873

Reviewed by
Julia Ive
Queen Mary University of London

Natural Language Processing (NLP) is a discipline at the crossroads of Artificial Intelligence (Machine Learning [ML] as its part), Linguistics, Cognitive Science, and Computer Science that enables machines to analyze and generate natural language data. The multi-disciplinary nature of NLP attracts specialists of various backgrounds, mostly with the knowledge of Linguistics and ML. As the discipline is largely practice-oriented, traditionally NLP textbooks are focused on concrete tasks and tend to elaborate on the linguistic peculiarities of ML approaches to NLP. They also often introduce predominantly either traditional ML or deep learning methods. This textbook introduces NLP from the ML standpoint, elaborating on fundamental approaches and algorithms used in the field such as statistical and deep learning models, generative and discriminative models, supervised and unsupervised models, and so on. In spite of the density of the material, the book is very easy to follow. The complexity of the introduced topics is built up gradually with references to previously introduced concepts while relying on a carefully observed unified notation system. The textbook is oriented to prepare the final-year undergraduate, as well as graduate students of relevant disciplines, for the NLP course and stimulate related research activities. Considering the comprehensiveness of the topics covered in an accessible way, the textbook is also suitable for NLP engineers, non-ML specialists, and a broad range of readers interested in the topic.

The book comprises 18 chapters organized in three parts. Part I, "Basics," discusses the fundamental ML and NLP concepts necessary for further comprehension using the example of classification tasks. Part II, "Structures," covers the principles of mathematical modeling for structured prediction, namely, for such structures as sequences and trees. Part III, "Deep Learning," describes the basics of deep learning modeling for classification and structured prediction tasks. The part ends with the basics of sequence-to-sequence modeling. The textbook thus emphasizes the close connection and inheritance between the traditional and deep-learning methods.

Following clear logic, generative models are introduced before discriminative ones (e.g., Chapter 7 from Part II introduces generative sequence labeling and Chapter 8 introduces discriminative sequence labeling), while modeling with hidden variables is presented at the end. Within each chapter, model descriptions are followed by their training and inference details. Finally, chapters are concluded with a summary, chapter notes, and exercises. The exercises are carefully designed to not only support and deepen comprehension but also stimulate further independent investigation on

https://doi.org/10.1162/coli_r_00423

© 2021 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International
(CC BY-NC-ND 4.0) license

the topic. Some example questions include: advantages of the variational dropout versus naïve dropout, or how sequence-to-sequence models could be used for sequence labeling.

In the following paragraphs I will cover the content of each chapter in more detail.

Chapter 1 opens Part I and gives an introduction to NLP and its tasks, as well as the motivation to focus on the ML techniques behind them as a uniting factor.

Chapter 2 gives a primer on probabilistic modeling, generative classification models, and Maximum Likelihood Estimation (MLE) training.

Chapter 3 explains representing documents with high-dimensional vectors, their clustering and classification with discriminative models, for both binary and multi-class settings using non-probabilistic algorithms (SVMs and Perceptron). The explanation of the differences between the generative and discriminative models using the concept of overlapping features is very helpful and succinct.

Chapter 4 focuses on the log-linear model and introduces probabilistic discriminative linear classifiers. The chapter details the training procedure with Stochastic Gradient Descent (SGD), as well as a range of popular loss functions.

Chapter 5 explains the basics of probabilistic modeling from the perspective of Information Theory and introduces such fundamental concepts as entropy, cross-entropy, and KL-divergence.

Chapter 6 introduces learning with hidden variables; more specifically, it focuses on the Expectation-Maximization (EM) algorithm and its applications in NLP (e.g., IBM Model 1 for Machine Translation). The algorithm is introduced gradually based on the earlier introduced k-means clustering.

Overall, I found the chapters of Part I to be a very good succinct introduction to a complex set of concepts.

Chapter 7 opens Part II and provides a guide into generative supervised (using MLE) and unsupervised (Baum-Welch algorithm) structured prediction using Hidden Markov Models (HMMs) for the sequence labeling task. It also explains such key algorithms as the Viterbi algorithm for decoding and the Forward-Backward Algorithm for marginal probabilities.

Chapter 8 complements Chapter 7 with the description of discriminative models for the sequence labeling task. The chapter starts with the local maximum entropy Markov models (MEMMs) and then provides the details of Conditional Random Fields (CRFs) with global training, as well as structured versions of Perceptron and SVMs. Chapter 9 explains how the previously seen sequence labeling models could be adapted to the sequence segmentation tasks using semi-Markov CRFs and beam search with Perceptron training.

Chapters 10 and 11 explain structured prediction for syntactic trees in constituent and dependency parsing. Chapter 10 starts with the generative MLE approach for constituent parsing, describing the CKY algorithm for decoding and the inside-outside algorithm for marginal probabilities. Then discriminative models and reranking are elaborated on.

Chapter 11 presents the advantages of addressing the structured prediction task (again for parsing) with transition-based modeling using non-local features and explains the shift-reduce constituent parsing.

Chapter 12 introduces the basics of Bayesian networks with the emphasis on training and inference methods building on previously introduced concepts, such as conditional independence and MLE. Bayesian learning is illustrated with the Latent Dirichlet Allocation (LDA) topic model and the Bayesian IBM Model 1 for alignment in statistical Machine Translation. I found this chapter to be slightly detached from the rest of the chapters in Part II and wish the Bayesian learning had been illustrated with any of the structured prediction models introduced earlier in this part (e.g., HMMs).

Chapter 13 opens Part III and introduces the basics of neural modeling for text classification, deriving from the generalized Perceptron model. Dense low-dimensional feature representations are presented as the main paradigm shift.

Chapters 14 and 15 follow the logic of Part II. Chapter 14 explains how neural network structures could be used to represent sequences of natural language. The chapter gives an in-depth overview of Recurrent Neural Networks (RNNs) as the principal architecture to represent sequences, as well as of an overview of attention mechanisms. Then, a range of networks is described that could be used to represent trees (e.g., Child-Sum Tree Long Short Term Memory networks [LSTMs]) and graphs (e.g., Graph Recurrent Neural Networks [GRNs]). The chapter is finalized with useful practical suggestions on how to analyze representations.

Chapter 15 details how local and global graph- and transition-based models for sequence labeling and parsing tasks could be built using neural networks. For a reader already familiar with neural networks, I suggest reading those chapters right after Chapter 11.

Chapter 16 elaborates on working with text in the input and output. The chapter starts with sequence-to-sequence modeling using LSTMs, then the authors show how this architecture could be augmented with attention and copying mechanisms. Finally, sequence-to-sequence models based on self-attention networks are explained. The last section of the chapter addresses the topical issue of semantic matching and presents a range of relevant models: Siamese networks, attention matching networks, and so forth.

Chapter 17 provides insights into pre-training and transfer learning. The chapter gives the details of neural language modeling and presents the non-contextualized embedding techniques as a by-product of this modeling, which I found to be very intuitive. Contextualized word embeddings are progressively introduced, starting with the RNN-based ones. Then the authors present how these embeddings could be built with self-attention networks. The chapter also explains such transfer learning techniques as pre-training, multitask learning, choice of parameters for sharing, and so on.

Chapter 18 culminates the book, focusing on deep learning with latent (hidden) variables. It opens with the modeling using categorical and structured latent variables, followed by the introduction of approximate variational inference for continuous latent variables. The chapter builds up on the knowledge of the EM algorithm. Particular attention is also paid to Variational Autoencoders (VAEs) and their usage for topic and text modeling.

In summary, this textbook provides a valuable introduction to Machine Learning approaches and methods applied in Natural Language Processing across paradigms. I strongly recommended it not only to students and NLP engineers, but also to a wider audience of specialists interested in NLP.

Julia Ive is a Lecturer in Natural Language Processing at Queen Mary University of London, UK. She is the author of many mono- and multimodal text generation approaches in Machine Translation and Summarization. Currently, she is working on the theoretical aspects of style preservation and privacy-safety in artificial text generation. Julia's e-mail address is `j.ive@qmul.ac.uk`.