Natural Language Processing

Jacob Eisenstein

Direct Studies - Book Report

Ryerson University

By Mahtab Tamannaee
Winter 2021

Introduction on this book...

Ryerson University

This book...

- Focus is on a core of the NLP, and the concepts of learning and search.
- **NLP problems** can be solved by a set of learning and search **methods**.
- How these **methods** work, and can be applied to NLP tasks.

- NLP Task Examples:
  - Document classification, word sense disambiguation, part-of-speech tagging, named entity recognition, parsing, coreference resolution, relation extraction, discourse analysis, language modeling, and machine translation.

This book: NLP Task Examples

- NLP Task Examples:
  - Document classification
  - Word sense disambiguation
  - Part-of-speech tagging
  - Named entity recognition
  - Parsing
  - Coreference resolution
  - Relation extraction
  - Discourse analysis
  - Language modeling
  - Machine translation.

Ryerson University

This book : Search and Learning Methods

- **Search Methods** : Viterbi, CKY, minimum spanning tree, shift-reduce, integer linear programming, beam search.

- **Learning Methods** : Maximum-likelihood estimation, logistic regression, perceptron, expectation maximization, matrix factorization, backpropagation.

The Organization of this book…

This book : Organization

This textbook is organized into **four** main units:

- Learning
- Sequences and trees.
- Meaning
- Applications

Ryerson University

This book : 1. Learning Unit

- This section builds up **a set of machine learning tools** that will be used in the other sections.

- Because the focus is on machine learning tools, *the text representations and linguistic phenomena are mostly simple:* "bag-of-words" text classification is treated as a model example.

- **Chapter 4** describes some of the more linguistically interesting applications of **word-based text analysis.**

Ryerson University

This book : 2. Sequences and trees Unit

- This section introduces the treatment of **language as a structured phenomena**.
- It describes **sequence and tree representations** and the algorithms that they facilitate, as well as the limitations that these representations impose.
- Chapter 9 introduces finite state automata and briefly overviews a context-free account of English syntax.

This book : 3. Meaning Unit

- This section takes a broad view of efforts to **represent and compute meaning from text**, ranging from ***formal logic*** to ***neural word embeddings***.

- It also includes two topics that are closely related to **semantics**: resolution of **ambiguous references**, and **analysis of multi-sentence structure**.

Ryerson University

This book : 4. Applications Unit

- The most prominent applications of NLP will be discussed in last chapters:
  - **1. Information Extraction**
  - **2. Machine Translation**
  - **3. Text Generation**

Ryerson University

The Chapters in this book…

This book:
Base NLP Chapters

- The review of probability in Appendix A
- Chapters 1-3 provide building blocks that will be used throughout the book
- Chapter 4 describes some critical aspects of the practice of language technology.
- Language models (chapter 6), sequence labeling (chapter 7), and parsing (chapter 10 and 11) are canonical topics in NLP distributed word embeddings (chapter 14)
  - Of the applications, machine translation (chapter 18) is the best choice: it is more cohesive than information extraction, and more mature than text generation.

Ryerson University

This book:
Machine Learning Chapters

- The chapter on unsupervised learning (chapter 5).
- The chapters on predicate-argument semantics (chapter 13), reference resolution (chapter 15), and text generation (chapter 19) are particularly influenced by recent progress in machine learning, including deep neural networks and learning to search.

This book:
Linguistic Orientation Chapters

The chapters on applications of sequence labeling (chapter 8), formal language theory (chapter 9), semantics (chapter 12 and 13), and discourse (chapter 16).

This book:
Application Chapters

The chapters on applications of sequence labeling (chapter 8), predicate-argument semantics (chapter 13), information extraction (chapter 17), and text generation (chapter 19).

*The **Notations** used in this book...*

As a general rule...

- Words, word counts, and other types of observations are indicated with **Roman letters (a, b, c)**.
- Parameters are indicated with **Greek letters (α, β, θ)**.
- Vectors are indicated with bold script for both **random variables x** and **parameters θ**.

Basic Notations :

Basics	
$\exp x$	the base-2 exponent, 2^x
$\log x$	the base-2 logarithm, $\log_2 x$
$\{x_n\}_{n=1}^N$	the set $\{x_1, x_2, \ldots, x_N\}$
x_i^j	x_i raised to the power j
$x_i^{(j)}$	indexing by both i and j

Ryerson University

Linear Algebra Notations :

Linear algebra	
$\boldsymbol{x}^{(i)}$	a column vector of feature counts for instance i, often word counts
$\boldsymbol{x}_{j:k}$	elements j through k (inclusive) of a vector $\boldsymbol{x}$
$[\boldsymbol{x}; \boldsymbol{y}]$	vertical concatenation of two column vectors
$[\boldsymbol{x}, \boldsymbol{y}]$	horizontal concatenation of two column vectors
$\boldsymbol{e}_n$	a "one-hot" vector with a value of 1 at position n, and zero everywhere else
$\boldsymbol{\theta}^\top$	the transpose of a column vector $\boldsymbol{\theta}$
$\boldsymbol{\theta} \cdot \boldsymbol{x}^{(i)}$	the dot product $\sum_{j=1}^{N} \theta_j \times x_j^{(i)}$
$\mathbf{X}$	a matrix
$x_{i,j}$	row i, column j of matrix $\mathbf{X}$
$\text{Diag}(\boldsymbol{x})$	a matrix with $\boldsymbol{x}$ on the diagonal, e.g., $\begin{pmatrix} x_1 & 0 & 0 \\ 0 & x_2 & 0 \\ 0 & 0 & x_3 \end{pmatrix}$
$\mathbf{X}^{-1}$	the inverse of matrix $\mathbf{X}$

Text Dataset Notations :

Text datasets	
w_m	word token at position m
N	number of training instances
M	length of a sequence (of words or tags)
V	number of words in vocabulary
$y^{(i)}$	the true label for instance i
$\hat{y}$	a predicted label
$\mathcal{Y}$	the set of all possible labels
K	number of possible labels $K = \lvert\mathcal{Y}\rvert$
$\square$	the start token
$\blacksquare$	the stop token
$\boldsymbol{y}^{(i)}$	a structured label for instance i, such as a tag sequence
$\mathcal{Y}(\boldsymbol{w})$	the set of possible labelings for the word sequence $\boldsymbol{w}$
$\Diamond$	the start tag
$\blacklozenge$	the stop tag

Ryerson Universi

Probability Notations :

Probabilities	
$\Pr(A)$	probability of event A
$\Pr(A \mid B)$	probability of event A, conditioned on event B
$\mathrm{p}_B(b)$	the marginal probability of random variable B taking value b; written $\mathrm{p}(b)$ when the choice of random variable is clear from context
$\mathrm{p}_{B\mid A}(b \mid a)$	the probability of random variable B taking value b, conditioned on A taking value a; written $\mathrm{p}(b \mid a)$ when clear from context
$A \sim p$	the random variable A is distributed according to distribution p. For example, $X \sim \mathcal{N}(0, 1)$ states that the random variable X is drawn from a normal distribution with zero mean and unit variance.
$A \mid B \sim p$	conditioned on the random variable B, A is distributed according to p.[2]

Ryerson University

Machine Learning Notations :

Machine learning	
$\Psi(\boldsymbol{x}^{(i)}, y)$	the score for assigning label y to instance i
$\boldsymbol{f}(\boldsymbol{x}^{(i)}, y)$	the feature vector for instance i with label y
$\boldsymbol{\theta}$	a (column) vector of weights
$\ell^{(i)}$	loss on an individual instance i
L	objective function for an entire dataset
$\mathcal{L}$	log-likelihood of a dataset
λ	the amount of regularization

Ryerson University