CS 6501 Natural Language Processing

Sequence Labeling (I)

Yangfeng Ji

September 19, 2019

Department of Computer Science
University of Virginia

Overview

1. Final Project Suggestion
2. Problem Definition
3. Hidden Markov Models
4. Parameter Estimation
5. Viterbi Decoding

Final Project Suggestion

Text Classification and Sentiment Analysis

Projects from last year:

- A Deeper Look into Social Media-focused Sentiment Analysis
- Citation Recommendation by Abstract

Convation Modeling

Projects from last year:

- Neural Dialog System with Personality
- Chit-chat Bot Modeling

Text Generaiton

Projects from last year:

- A Deep Learning Approach for Meme Generation
- Variational Image Captioning Using Deterministic Attention
- Neural Style Transfer for Natural Language
- Generating Subject Line from Email Text

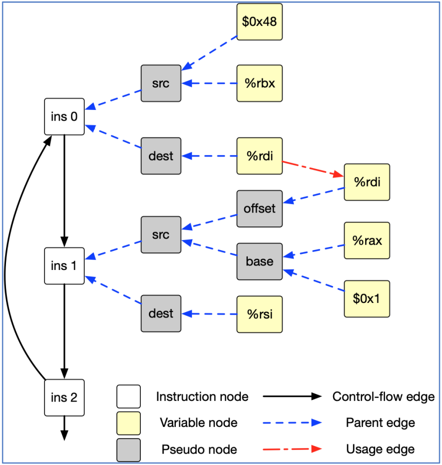
Interpretability and Adversarial Learning

Projects from last year:

- Incorporating Textual Data from Reviews for Explainable Recommendation
- Label Flipping Attacks on Sentiment Analysis Systems
- Topic-based Interpretability Improvement on Citation Recommendation System

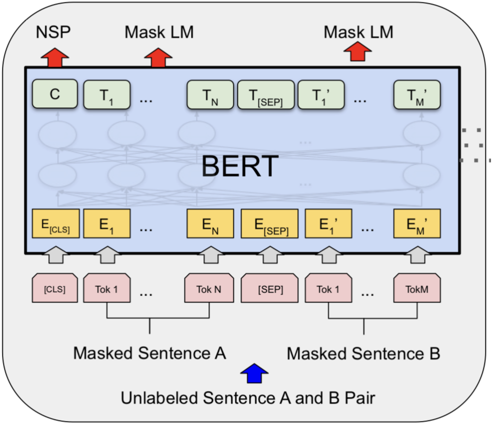
Sequential Modeling for Computer Architecture

- Using neural network and NLP models to analyze machine code



Ashish Venkat
(venkat@virginia.edu)

Running BERT on Edge Devices



Marzieh Lenjani
(marzieh.lenjani@gmail.com)

Problem Definition

Part of Speech

- A way to categorize words with similar *grammatical* properties
- Common English POS tags
  - NOUN: used to name persons, things, animals, places etc.
    e.g., `Tom Hanks`, `yesterday`, `Grounds`
  - VERB: show an action or state
    e.g., `fight`, `was`
  - PRONOUN: replacement of nouns
    e.g., `she`, `his`, `it`, `theirs`
  - ADJECTIVE: used to describe a noun or a pronoun
    e.g., `large`, `beautiful`

Part of Speech (II)

- Common English POS tags (cont.)
  - ADVERB: used to describe adjectives, verbs, or another adverb
    e.g., `gracefully`, `yesterday`, `very`
  - PREPOSITION: specify location or a location in time
    e.g., `above`, `near`, `since`
  - CONJUNCTION: join words, phrases, or clauses together
    e.g., `and`, `for`
  - INTERJECTION: convey strong emotions
    e.g., `Ouch`, `Hey`

POS Tagging

Example

Teacher Strikes Idle Children

- Teacher$_{\texttt{Noun}}$ Strikes$_{\texttt{Noun}}$ Idle$_{\texttt{Verb}}$ Children$_{\texttt{Noun}}$
- Teacher$_{\texttt{Noun}}$ Strikes$_{\texttt{Verb}}$ Idle$_{\texttt{Adj}}$ Children$_{\texttt{Noun}}$

[Eisenstein, 2018, Chap 8]

Sequence Labeling

From a training set, to learn a mapping $p(\boldsymbol{y} \mid \boldsymbol{x})$,

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(\boldsymbol{y} \mid \boldsymbol{x}) \tag{1}$$

where

- $\boldsymbol{x}$: a sentence (a sequence of words)
- $\boldsymbol{y}$: the POS tag sequence of $\boldsymbol{x}$ (a sequence of POS tags)

Example

$\boldsymbol{x}$	x_1 Teacher	x_2 Strikes	x_3 Idle	x_4 Children
$\boldsymbol{y}$	y_1 NOUN	y_2 NOUN	y_3 VERB	y_4 NOUN

Label Classification

$$\hat{y} = \operatorname*{argmax}_{y} p(\boldsymbol{y} \mid \boldsymbol{x}) \tag{2}$$

- $\boldsymbol{x}$: entire sentenceonly one token
- $\boldsymbol{y}$: entire sequenceonly the corresponding tag
- $P(\boldsymbol{y}|\boldsymbol{x})P(y_i|x_i)$

Example

$\boldsymbol{x}$	x_1 Teacher	x_2 Strikes	x_3 Idle	x_4 Children
$\boldsymbol{y}$	y_1 NOUN	y_2 NOUN	y_3 VERB	y_4 NOUN

Label Classification

Example

$\boldsymbol{x}$	x_1	x_2	x_3	x_4
	Teacher	Strikes	Idle	Children
$\boldsymbol{y}$	y_1	y_2	y_3	y_4
	NOUN	NOUN	VERB	NOUN
	NOUN	VERB	ADJ	NOUN

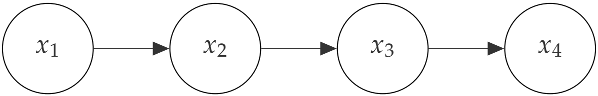
Limitations [`TODO: revise the following description`]

- No constraint from the previous POS tag
  - Solution: sequence labeling (e.g., hidden Markove models, conditional random fields)
- No information from the surrounding words
  - Solution: conditional random fields

Markov Chain

Modeling the dependency between $\{y_i\}$ as

$$p(\boldsymbol{y}) = p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_2)p(y_4 \mid y_3) \tag{3}$$



Question

How to merge the conditional dependence from $p(\boldsymbol{y})$ into $p(\boldsymbol{y} \mid \boldsymbol{x})$?

Hidden Markov Models

Generative Models

- Observation $\boldsymbol{x}$
- Target variable $\boldsymbol{y}$

Bayes rule

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\boldsymbol{y}) \cdot p(\boldsymbol{y})}{p(\boldsymbol{x})} \\ &\approx \underbrace{p(\boldsymbol{y})}_{\text{prior}} \cdot \underbrace{p(\boldsymbol{x}|\boldsymbol{y})}_{\text{likelihood}} \end{aligned} \tag{4}$$

Prior $p(\boldsymbol{y})$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) \approx p(\boldsymbol{y}) \cdot p(\boldsymbol{x} \mid \boldsymbol{y}) \tag{5}$$

Factorization

$$p(\boldsymbol{y}) = \prod_{i=1} \underbrace{p(y_i \mid y_{i-1})}_{\text{Transition probability}} \tag{6}$$

Graphical model

$$y_1 \longrightarrow y_2 \longrightarrow y_3 \longrightarrow y_4$$

$$p(\boldsymbol{y}) = p(y_1) \cdot p(y_2 \mid y_1) \cdot (y_3 \mid y_2) \cdot p(y_4 \mid y_3)$$

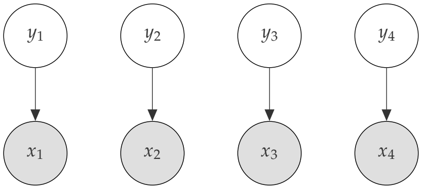
Likelihood $P(x|y)$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = p(\boldsymbol{y}) \cdot p(\boldsymbol{x} \mid \boldsymbol{y}) \tag{7}$$

Factorization

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \prod_{i=1} \underbrace{p(x_i \mid y_i)}_{\text{Emission probability}} \tag{8}$$

Graphical model

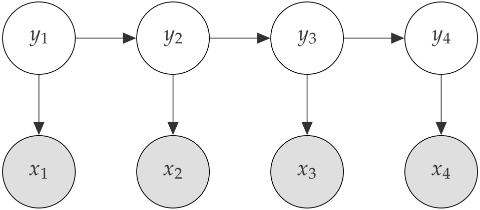


$$p(\boldsymbol{x} \mid \boldsymbol{y}) = p(x_1 \mid y_1) \cdot p(x_2 \mid y_2) \cdot p(x_3 \mid y_3) \cdot p(x_4 \mid y_4)$$

Hidden Markov Models

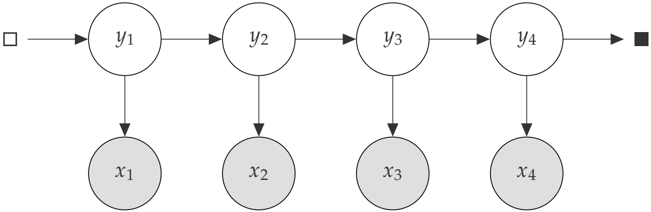
$$P(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1} \left\{ P(y_i | y_{i-1}) P(x_i | y_i) \right\} \tag{9}$$

Graphical model



- $\boldsymbol{x}$: observation (e.g., sentences)
- $\boldsymbol{y}$: hidden variables (e.g., POS sequences)

Two Special Tokens



$$P(\boldsymbol{x}, \boldsymbol{y}) = P(y_1 \mid \square) \prod_{i=1} \left\{ P(y_i | y_{i-1}) P(x_i | y_i) \right\} p(\blacksquare \mid y_4) \quad (10)$$

Two Questions

- Learning: parameter estimation
  - $p(y_n \mid y_{n-1}) =?$
  - $p(x_n \mid y_n) =?$
- Prediction: inference/decoding
  - $\hat{\boldsymbol{y}} = \operatorname{argmax}_{\boldsymbol{y}} p(\boldsymbol{y} \mid \boldsymbol{x})$

Parameter Estimation

Training Corpus

Training corpus

- `they`$_{\text{PRON}}$ `can`$_{\text{VERB}}$ `fish`$_{\text{NOUN}}$
- `teacher`$_{\text{NOUN}}$ `strikes`$_{\text{VERB}}$ `idle`$_{\text{ADJ}}$ `children`$_{\text{NOUN}}$
- ...

How to estimate the following probabilities?

$$\begin{aligned} P(x_i|y_i) &= ? \\ P(y_i|y_{i-1}) &= ? \end{aligned} \tag{11}$$

MLE

Transition probability

$$P(y_i|y_{i-1}) = \frac{\#(y_i, y_{i-1})}{\#(y_{i-1})} \tag{12}$$

Emission probability

$$P(x_i|y_i) = \frac{\#(x_i, y_i)}{\#(y_i)} \tag{13}$$

Viterbi Decoding

Decoding by Brute-force Search

Given a sentence

```
The dog barks
```

and the possible POS tags {D,N,V}. A brute-force algorithm will try every possible combination as

$$\begin{array}{ccc} D & D & D \\ D & D & N \\ D & D & V \\ D & N & D \\ D & N & N \\ D & N & N \\ \vdots & \vdots & \vdots \end{array}$$

There are $3^3 = 27$ possible sequences in this case.

Decoding y_i

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}} p(\boldsymbol{y} \mid \boldsymbol{x}) \tag{14}$$

$$= \operatorname*{argmax}_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}) \tag{15}$$

$$\tag{16}$$

With conditional dependency

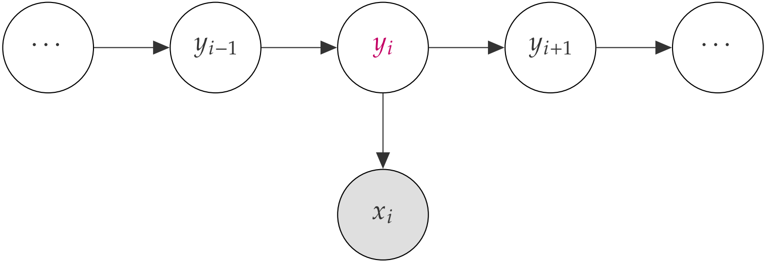
$$p(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1} \left\{ p(y_i \mid y_{i-1}) p(x_i \mid y_i) \right\} \tag{17}$$

$$= \cdots \underbrace{p(y_i \mid y_{i-1}) \cdot p(y_{i+1} \mid y_i) \cdot p(x_i \mid y_i)}_{\text{items related to } y_i} \cdots \tag{18}$$

Decoding y_i (II)

$$p(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1} \left\{ p(y_i \mid y_{i-1}) p(x_i \mid y_i) \right\} \tag{19}$$

$$= \cdots \underbrace{p(y_i \mid y_{i-1}) \cdot p(y_{i+1} \mid y_i) \cdot p(x_i \mid y_i)}_{\text{items related to } y_i} \cdots \tag{20}$$



Factorization

Factorize $p(\boldsymbol{x}, \boldsymbol{y})$ with respect to (x_i, y_i)

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{y}) &= p(\boldsymbol{x}_{\leq i-1}, \boldsymbol{y}_{\leq i-1}) \cdot p(x_i, y_i | \boldsymbol{y}_{\leq i-1}) \cdot p(\boldsymbol{x}_{\geq i+1}, \boldsymbol{y}_{\geq i+1} | y_i) \\
&= p(\boldsymbol{x}_{\leq i-1}, \boldsymbol{y}_{\leq i-1}) \cdot p(x_i | y_i) \cdot p(y_i | y_{i-1}) \qquad (21) \\
&\quad \cdot p(\boldsymbol{x}_{\geq i+1}, \boldsymbol{y}_{\geq i+1} | y_i) \qquad (22)
\end{aligned}
$$

Three components

$$
\underbrace{p(\boldsymbol{x}_{\leq i-1}, \boldsymbol{y}_{\leq i-1})}_{\leq i-1} \cdot \underbrace{p(x_i | y_i) \cdot p(y_i | y_{i-1})}_{i} \cdot \underbrace{p(\boldsymbol{x}_{\geq i+1}, \boldsymbol{y}_{\geq i+1} | y_i)}_{> i} \qquad (23)
$$

Principle of Optimality

Assume we have the optimal decoded sequence $\hat{\boldsymbol{y}}$, such that

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}), \tag{24}$$

then, $\boldsymbol{y}_{\geq i}$ is the also the optimal subsequence of the following subproblem

$$\hat{\boldsymbol{y}}_{\geq i} = \operatorname*{argmax}_{\boldsymbol{y}_{\geq i}} p(\boldsymbol{x}_{\geq i}, \boldsymbol{y}_{\geq i} \mid \hat{y}_i) \tag{25}$$

Justification:

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}) = p(&\boldsymbol{x}_{\leq i-1}, \hat{\boldsymbol{y}}_{\leq i-1}) \\ &\cdot p(x_i \mid y_i) \cdot p(y_i \mid \hat{y}_{i-1}) \\ &\cdot p(\boldsymbol{x}_{\geq i+1}, \boldsymbol{y}_{\geq i+1} | y_i) \end{aligned} \tag{26}$$

Basic Idea of Decoding

$$\underbrace{p(\boldsymbol{x}_{\leq i-1}, \boldsymbol{y}_{\leq i-1})}_{\leq i-1} \cdot \underbrace{p(x_i|y_i) \cdot p(y_i|y_{i-1})}_{i} \cdot \underbrace{p(\boldsymbol{x}_{\geq i+1}, \boldsymbol{y}_{\geq i+1}|y_i)}_{>i} \tag{27}$$

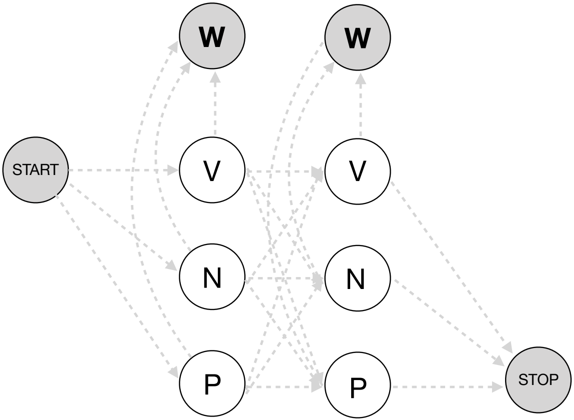
- Forward computation: start from y_1, for every possible value of y_i, from the best path from y_{i-1}

  $$\max_{y_{i-1}} p(\boldsymbol{x}_{\leq i-1}, \boldsymbol{y}_{\leq i-1}) \cdot p(x_i|y_i) \cdot p(y_i|y_{i-1}),$$

  depending on past and present states $\{\boldsymbol{y}_{\leq i}\}$
- Backward tracing: start from $y_T = \blacksquare$, for a given y_{i+1} find the best y_i

Example

$$\max_{y_{i-1}} p(\boldsymbol{x}_{\leq i-1}, \boldsymbol{y}_{\leq i-1}) \cdot p(x_i|y_i) \cdot p(y_i|y_{i-1}),$$



Reference

Eisenstein, J. (2018).
Natural Language Processing.
MIT Press.