CS 378 Lecture 8

Today

- Bias in embeddings
- Part-of-speech intro
- Sequence tagging
- Tagging with classifiers

lexical semantics
300-d vector space

Announcements   [prons in Zoom]

- A1
- A2
- Survey:
    Good: rewatchable lectures, exercises/breakouts, notes
  Not so good: ① Working in pairs on HW - please do!
               ② More exercises
               ③ Bigger breakouts/more engagement
               ④ More context/big picture
               ⑤ Reading guidance — bolding

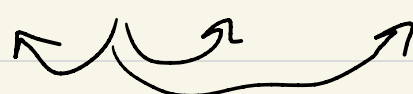# Recap Skip-gram

Input: corpus of text

Corpus $\Rightarrow (x, y)$ pairs which are $\leq k$ words apart (k = window size)

$$P(\text{context} = y \mid \text{word} = x) = \frac{e^{\overline{v_x} \cdot \overline{c_y}}}{\sum\limits_{y' \in V} e^{\overline{v_x} \cdot \overline{c_{y'}}}}$$

Maximize log likelihood of data $\Rightarrow$ get useful $\overline{v_s}, \overline{c_s}$

Use $\overline{v}, \overline{c}$, or $\overline{v} + \overline{c}$ in downstream tasks

the fish swam quickly $k = 2$

Where we are

Classification: $\text{argmax}_{Y} \; \overline{w}_Y^T f(\overline{x})$

$\underbrace{}_{\text{NN, BoW}}$

sent

doc $\Rightarrow$ label

sent $\Rightarrow$ label for each word in that sentence

Part-of-speech tagging

① Structurally different problem

$\overline{X} \rightarrow y$

$X_1, \dots, X_n \rightarrow Y_1, \dots, Y_n$ tag for each word

② Syntax

Parts-of-speech

Text to speech: <u>record</u>

Info. extraction: <u>arms</u>   V or N?

POS Tags

Open — class: there can be new words here	Closed — class: fixed set
Nouns: Proper: IBM, Common: Cat	Determiners: the, a (articles)
Verbs: see, registered Adjs: yellow Advs: swiftly	Some DT + N ⇒ NP Conjunctions: and/or Pronouns Prepositions: up, on, ... Particles: made <u>up</u>

Auxiliaries, modals: Aux: had [V]
 Modals: could / would / should

"real" Penn Treebank

fed — NNP proper noun — I fed her...
VBD past tense verb
VBN participial — I had fed...

raises — NNS plural noun
VBZ 3rd person present verb

interest — NN noun
VBP — I interest you
VB infinitive: I want to interest you

rates — NNS
~~VBZ~~

0.5 — CD cardinal

percent — NN

(yellow) correct   (teal) rates are a living thing that get interested

Sequence Tagging

Input: $\bar{x} = (x_1, ..., x_n)$ $x_i \in V$ words

Output: $\bar{y} = (y_1, ..., y_n)$ $y_i \in \mathcal{Y}$ tags

Structured classification: output has
structure

Start: use classifiers

$P(y_i = y \mid \bar{x})$ use LR, .. to assign
tag y to word i in
sentence $\bar{x}$

Next: Hidden Markov Models