

CS378 Lecture 9: Sequence Labeling, POS, HMMs

Announcements

- A2 due Tues
- Bias in embs response
- A1 back soon

Survey

- Lower Zoom volume
- Exercises / pace
- Ling / language
- More state-of-the-art

Recap Bias: $\text{sim}(\text{she}, \text{receptionist}) > \text{sim}(\text{he}, \text{receptionist})$

Multilingual: dictionaries can help us align embeddings across languages

Today Sequence labeling

POS: definitions

Why sequence models for this?

HMM intro (if time)

Where we are

Classification: $\operatorname{argmax}_y \bar{w}_y^T f(\bar{x})$

\bar{x}

NN or bag of words

~ sent
or

y
→ label

binary/multi

 doc

Tagging =

$y_1 \ y_2 \ y_3 \ y_4$
↑ ↑ ↑ ↑
~

label for each word
sent

Part-of-speech tagging

Input: sentence x_1, \dots, x_n

Output: POS labels y_1, \dots, y_n for each word

Why POS?

Teacher strikes idle kids

N	N	V
N	V	ADJ

disambiguate
meaning

Text-to-speech: record

POS Tags

Open-class tags: new words with these tags are always coming out

Closed-class tags: more like function words in English

Open-class (NNP)

(N) Nouns: Proper: Google

Common: cat, company
(NN, NNS)

(V) Verbs: see, registered

Adjectives: yellow

Adverbs: swiftly

Closed-class

Determiners: the, a (articles)

some, many

$DT + N \Rightarrow NP$

Conjunctions: and, or

Prepositions: up, on, in, to, ...

Particles: made up

Auxiliary verbs: had [V]

Modal verbs: could / would / should

Fed raises interest rates 0.5 percent

Fed **NNP** proper noun
 VBD past tense verb (I fed)
 VBN participial (I had fed)

raises **NNS** plural noun
 VBZ 3rd person present verb

interest **NN**
 VBP I interest you
 VB infinitive: to interest (I want you to interest me)

rates **NNS** **Correct**
 VBZ **alt**

0.5 **CD** cardinal **alt**

percent **NN** **Correct**

Fed raises interest rates:
 1 2 3 4

(2) Features that depend on i

$$f(\bar{x}, i)$$

$f(x_i) \leftarrow$ looks at one word, but

Look at words "around" position i

we want word in context

$$f(\bar{x}, i=3) = \begin{cases} \text{Prev Word} = \text{raises} \\ \text{Curr Word} = \text{interest} \\ \text{Next Word} = \text{rates} \end{cases}$$

$f(\bar{x}, i=4) \dots$

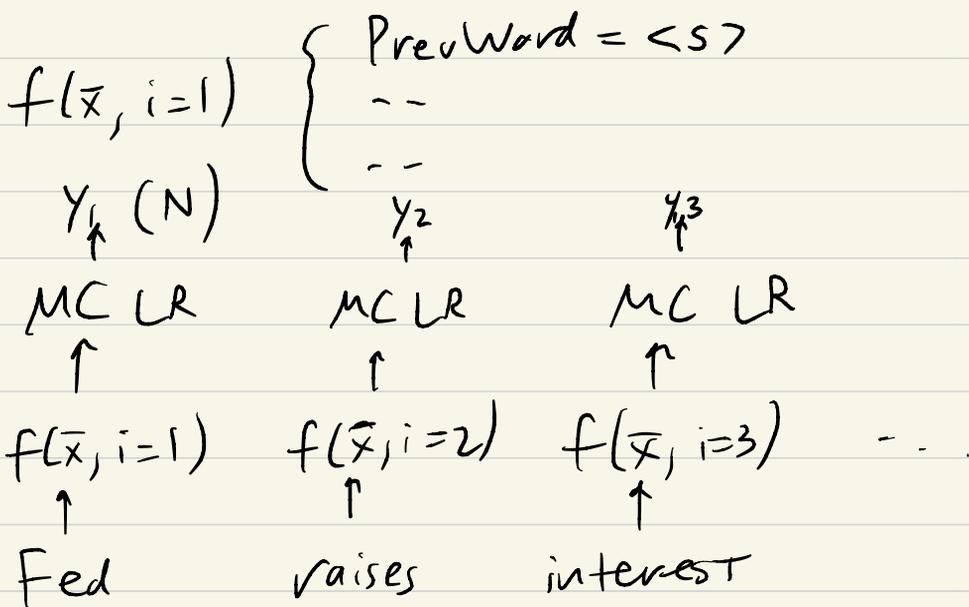
$$f(\bar{x}, i=3) = \begin{bmatrix} 0 & 0 & 1 \\ & & \text{Prev Word} = \text{raises} \\ & & | \\ & & \dots \\ & & \text{Curr Word} = \text{interest} \\ & & | \\ & & \dots \\ 0 & & 1 \end{bmatrix}$$

"bag of positional words"

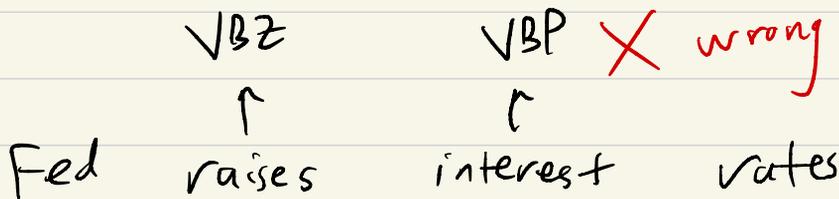
Curr Word = raises

3 - |V| feats

start-of-sentence



Problems with this



We know this is wrong. VBZ-VBP will rarely happen

We want to prohibit this... how?

We want to model sequences

$$P(\bar{y} | \bar{x}) \quad (\text{MC LR: } \prod_{i=1}^n P(y_i = y | \bar{x}))$$

↑
seq of labels

Structured prediction

predicting a sequence, tree, graph, ...

Hidden Markov Model: $P(\bar{x}, \bar{y})$

↑ ↑
words tags