

## Lecture 9 – Tagging Continued

### Goals

1. Language Model RNN
2. Discriminative Tagging
  - a. Max Entropy
  - b. Simple RNN
  - c. Bidirectional RNN
3. Decoding (or inference) with RNN
4. LSTM

Review of LM RNN

$$p_k^t = p(w_t = k | w_1, w_2, \dots, w_{t-1})$$

$$\Theta = \{W^{h,x}, W^{h,h}, W^{o,h}\}$$


---

New graphical representation

$$h^t = \tanh(W^{h,x} x^t + W^{h,h} h^{t-1})$$

$$p^t = \text{softmax}(W^{o,h} h^t)$$

Goal:  $\Theta = \underset{\Theta}{\operatorname{argmax}} \sum_{s \in S} \sum_{\substack{t \\ \text{pos in sentence}}} \log p_{w_t}^t$

$S = \text{Sentences}$

---

Discriminative Tagging

tags  $y_1, y_2, y_3, \dots$   
 words This is a ...  
 $w_1, w_2, w_3, \dots$

$$p(y_1, \dots, y_n | w_1, \dots, w_n) = \prod_{i=1}^n p(y_i | w_i, \dots, w_n, y_1, \dots, y_{i-1})$$

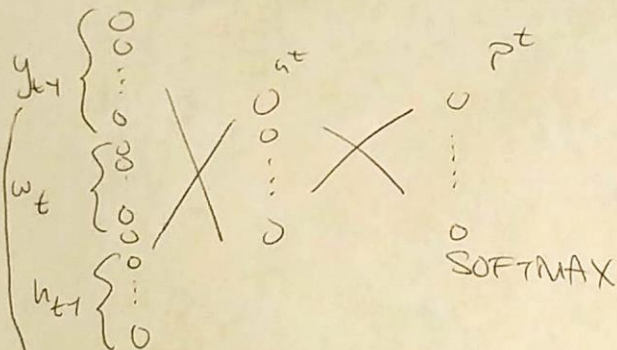
Max Entropy Model:

$$p(y | w) = \frac{1}{z(w, \theta)} e^{\theta f(w, y)}$$

Represent as feed forward NN

only connect with corresponding output tags

## Simple RNN



$$p_k^t = p(y_t = k | y_1, \dots, y_{t-1}, \underbrace{w_1, \dots, w_t}_{\text{not future words}})$$

Must keep  $y_{t-1}$  b/c :

Suppose we have

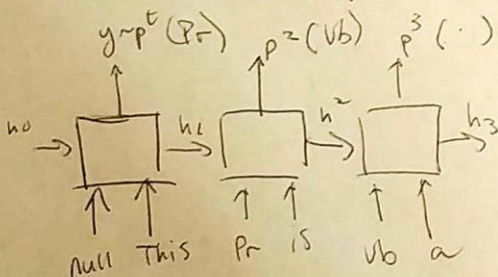
$y_1, y_1, y_1, y_1, \dots$   
a a a a

$h_{t-1}$  gives distribution over past choices (of tags)  
but not the actual  $y_{t-1}$ .

With one hot.

can't deal with unseen data

## Decoding with RNN



Sample from  $p(y_1, \dots, y_n | w_1, \dots, w_n)$

