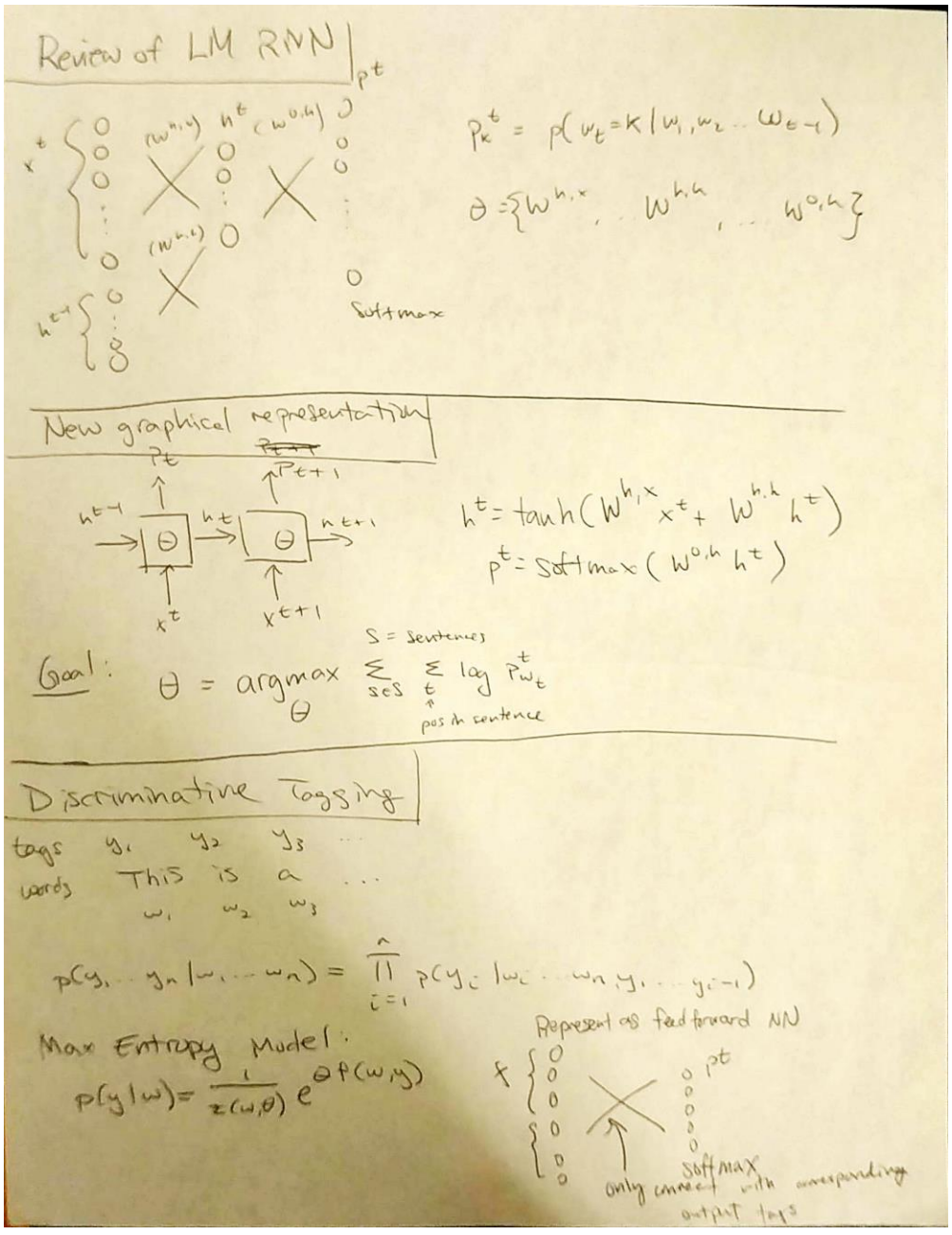
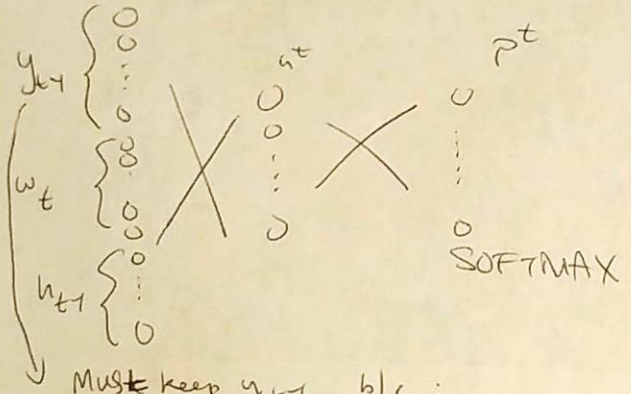
**Lecture 9 – Tagging Continued**

**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



## Simple RNN




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

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

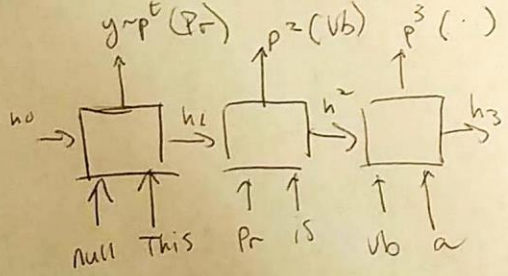
Suppose we have

$y_1$ $y_1$ $y_1$ $y_1$ ...
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 $p(y_1 \dots y_n \mid w_1, \dots w_n)$