Lecture 8 – Tagging

Goals

1. Generative Tagging
   a. Parameterization
   b. Estimation
   c. Inference
2. Discriminative tagging
   a. Max Entropy
   b. Simple RNN
   c. Bidirectional RNN

Generative Tagging

- Sentence: I love white dogs
- Tags: Pr, Vb Adj, N
- Baseline: most frequent tag – already gets 90%+ accuracy
- We are given
  - $S = w_1, \dots, w_n$
  - $T = y_1, \dots, y_n$
- Parameterization
  - $P(S,T) = p(w_1, \dots, w_n, y_1, \dots, y_n) = \prod p(w_i | w_1, \dots, w_n, y_1, \dots, y_i) p(y_i | w_1, \dots, w_n, y_1, \dots, y_{i-1})$
- Markov Assumption
  - $P(S,T) = \prod p(w_i | y_i) p(y_i | y_{i-1})$
    - $p(w_i | y_i)$ called the emission probability
    - $p(y_i | y_{i-1})$ called the transition probability
  - Number of parameters: $VT + T^2$
- Estimation
  - MLE of transition probability: $p(Vb|Pr) = \frac{count(Pr,Vb)}{count(Pr)}$
  - MLE of emission probability: $p(I|Pr) = \frac{count(I,Pr)}{count(Pr)}$
  - Remember there are smoothing issues
  - Want to use EM…but
- Inference Problem
  - We are looking for $T^* = argmax_T\big(P(S,T)\big)$
    - Assume $p(t_i | t_{i-1})$ and $p(w_i | t_i)$ are given
  - Solution: Viterbi Algorithm
    - Let n be the length of the sequence
    - $\pi[i,t] \rightarrow \max \log(probability\ of\ a\ sequence\ that\ ends\ at\ position\ i\ with\ tag\ t)$
    - Goal: $\max_{t \in T} \pi[n,t]$
    - Base Case: $\pi[0,*] = \log(1), \pi[0,t] = \log(0) = -\infty$

        - Recursive Case
            - $\pi[i,t] = \max_{t_{prev}} \pi[i_{prev}, t_{prev}] + \log\left(p\left(t \middle| t_{prev}\right)\right) + \log(p(i|t))$
        - Complexity: $O(nT^2)$
- Now we can do estimation with EM
    - Take a random guess of the parameters, and compute the MLE efficiently with the Viterbi algorithm
    - Bad results with random initialization
    - Can get good results with good initialization

Discriminative Tagging

- Now we estimate
    - $p(T|S) = p(y_1, \dots, y_n | w_1, \dots, w_n) = \prod_i p(y_i | w_1, \dots, w_n, y_1, \dots, y_{i-1})$
- Max Entropy
    - $p(y|w) = \frac{1}{z(\theta,w)} e^{\theta f(y,w)}$
        - Indicator function $f(y,w)$ of features $1 \dots K$
        - i.e. $f_1(y,w) = \begin{cases} 1, if\ y = noun, w\ is\ capitalized \\ 0, o.w. \end{cases}$
- NN Structure for Max Entropy
    - Final layer: softmax, each unit represents $p(y_t = k|w)$
    - Vector representing indicator function for y, w