

HMM Tagging Model

Jean Mark Gawron

March 18, 2009

1 HMM Tagging model

The HMM tagging model is the following:

$$P(w_{1,n}, t_{1,n}) = \prod_{i=1}^n P(w_i | t_i) * P(t_i | t_{i-1}) \quad (1)$$

This means take the product of all the

$$P(w_i | t_i) * P(t_i | t_{i-1})$$

Or, equivalently we can express this as a sum using log probabilities.

$$\log P(w_{1,n}, t_{1,n}) = \sum_{i=1}^n \log P(w_i | t_i) + \log P(t_i | t_{i-1}) \quad (2)$$

Why? Because:

$$a_1 \times a_2 \times \dots \times a_n = P \quad \text{iff} \quad \log a_1 + \log a_2 + \dots + \log a_n = \log P$$

2 Max Likelihood estimates

$$P(w | t) = \frac{\text{count}(w, t)}{\text{count}(t)}$$
$$P(t_i | t_{i-1}) = \frac{\sum_{w_j, w_k} \text{count}(w_j, t_{i-1}, w_k, t_i)}{\text{count}(t_{i-1})}$$

In brief, for

$$P(w | t)$$

Divide the number of times the word occurs with the tag by the number of occurrences of the tag.

For example,

$$P(\text{walked} \mid \text{VBN}) = \frac{\text{Count}(\text{walked_VBN})}{\text{Count}(\text{VBN})}$$

And for

$$P(t_i \mid t_{i-1})$$

Divide the number of times tag t_i follows tag t_{i-1} by the number of occurrences of tag t_{i-1} .

For example,

$$P(\text{NN} \mid \text{VBN}) = \frac{\sum_{w_i, w_j} \text{Count}(w_i\text{-VBN } w_j\text{-NN})}{\text{Count}(\text{NN})}$$