**Lecture 2 – Smoothing (Continued)**

1. Methods for smoothing
    a. Add $\alpha$ (see lecture 2)
    b. Linear Interpolation (see lecture 2)
    c. Discounting (Kneser-Ney Smoothing)
        i. Very successful in NLP
        ii. Idea: Take some probability mass from all existing bigrams, and redistribute (in some way, the probability among unseen bigrams)
            1. Create a new distribution $c^*(w_{i-1}, w_i)$
            2. $c^*(w_{i-1}, w_i) = \max(count(w_{i-1}, w_i) - d, 0), d > 0$
            3. $p(w_i, w_{i-1}) = \begin{cases} \frac{c^*(w_{i-1}, w_i)}{count(w_{i-1})}; & if\ count(w_{i-1}, w_i) > 0 \\ \alpha(w_{i-1}) \frac{p(w_i)}{\sum_{w \in \{w | count(w_i, w) = 0\}} p(w)}; otherwise \end{cases}$
        iii. Promiscuity
            1. Redistributing based on unigram probability not always good
            2. Instead, let $P_C(w)\alpha|\{w_{i-1}|count(w_{i-1}, w_i) > 0\}|$

**Lecture 3 – Topic Models and EM**

1. Syntax trees shown to help statistical NLP, but not neural NLP
2. Higher level models – semantics trees? Not convincingly helpful as well
3. How to model whole documents of text?
    a. Hierarchical Segmentation?
    b. Centering?
    c. RST?
4. Topic Models
    a. Choose topics in a document, and generate text from those topics
        i. Learned in an unsupervised manner
    b. Blend = topic distribution (specific to individual documents)
        i. $\theta_z \geq 0, \sum_z \theta_z = 1$
    c. Topic = distribution over words (shared across the collection of documents)
        i. $\beta_{w|z, w \in V} \geq 0, \sum_{w \in V} B_{w|z} = 1$
    d. Generally, some topics found make sense…but many (most?) do not
    e. Example
        i. $V = \{r, g, b\}$
        ii. Topic 1: $\beta_{r|1} = \beta_{g|1} = .5, \beta_{b|1} = 0$
        iii. Topic 2: $\beta_{r|2} = \beta_{g|2} = 0, \beta_{b|1} = 0$
        iv. $\theta_1 = \theta_2 = \frac{1}{2}$
    f. How to generate documents, given $\theta, \beta, n$?
        i. Model 1 (Not a topic model)
            1. $z \sim Categ(\theta_1, \theta_2)$
            2. For $i = 1 \ldots n$, $w_i \sim Categ(\beta_{w|z})$

ii. Model 2 (word order oblivious)
    1. For $i = 1 \dots n$, $z_i \sim Categ(\theta_1, \theta_2) \wedge w_i \sim Categ\left(\beta_{w|z_i}\right)$

g. How can we compute $p(w_1, \dots, w_n)$?
    i. Model 1
        1. $\sum_z \theta_z \prod_w \beta_{w|z}$
    ii. Model 2
        1. $\prod_w \sum_z \theta_z \beta_{w|z}$

h. How to estimate $\theta$ and $\beta$?
    i. Observed case
        1. Given $z_1, \dots, z_k \wedge w_1, \dots, w_n$
            a. $\hat{\theta}_z = \frac{count(z)}{n}$
            b. $\hat{\beta}_{w|z} = \frac{count(w,z)}{count(z)} = \frac{count(w,z)}{\sum_{w\prime} count(w\prime,z)}$
    ii. Unobserved case
        1. Use stochastic gradient descent
        2. Use EM algorithm (will discuss next time)