Lecture 4 – Expectation Maximization (EM)

1. Topic Models
2. Parameter estimation for TM using EM
3. EM for MT (IBM Model 1)

Clarification about Generative Models

- Remember we have θ_z (different for each document) and $\beta_{w|z}$ (same for all documents)
- Assume we have k topics and n words
- Option 1
  - $z \sim categ(\theta_1, \theta_2)$ controls the proportion of topics
  - For $i = 1 \dots n$, $w_i \sim categ(\beta_{w|z})$
- Option 2
  - For $i = 1 \dots n$, $z_i \sim categ(\theta_1, \theta_2), w_i \sim categ(\beta_{w|z_i})$

EM

- Extremely useful optimization procedure in NLP
- Consider the observed case
  - The observations are pairs $(z_1, w_1), (z_2, w_2), \dots$
  - We can estimate $\hat{\theta}_z = \frac{count(z)}{n} = \frac{\sum_{i=1}^{n}[[z_i = z]]}{n}$ and $\hat{\beta}_{w|z} = \frac{count(w,z)}{count(z)}$
- What do we do in the unobserved case?
  - We want to maximize $p(w_1, \dots, w_n) = \prod_{i=1}^{n}\sum_{z_i=1}^{k}\theta_{z_i}\beta_{w_i|k_i}$
  - We can use stochastic gradient descent or EM; today we discuss EM
  - For a given w, we don't know the z – z is unobserved.
    - Want to estimate $p(z_i = z|w_i, \beta, \theta) = \frac{p(z_i = z|\beta, \theta)}{\sum_{z'} p(z_i = z', w_i|\beta, \theta)} = \frac{p(z_i = z, w_i|\beta, \theta)}{p(w_i|\beta, \theta)} = \frac{\theta_z \beta_{w_i|z}}{\sum_{z'}\theta_{z'}\beta_{w_i|z'}} = q(z|i)$
    - This is the expectation step – given my parameters, what is the distribution of the hidden variable?
    - Now for the maximization step
    - $\hat{\theta}_z = \frac{\sum_{i=1}^{n} q(z|i)}{n}$, $\hat{\beta}_{w|z} = \frac{\sum_{i=1}^{n} q(z|i)[[w_i = w]]}{\sum_{i=1}^{n} q(z|i)}$

Application to MT

- "the blue house" → "la maran blue"
- Alignment $a = \{1, 3, 2\}$ (records where each French word was in the English phrase)
- Observed Case:
  - We assume every word is translated independently
  - $p(f|e, a) = \prod_{j=1}^{n} p(f_j|e_{a_j})$
  - i.e. $p(la|the) = \frac{count(la, the)}{count(the)}$

- Estimate the hidden variable a: $p(a|f,e) = \frac{p(a,f|e,\theta)}{\sum_{a'} p(a',f|e,\theta)}$
- Maximization: $p(a,f|e,\theta) = p(a|e,\theta)p(f|e,a,\theta) = \prod p(f_j|e_j,\theta)$
  - Assume $p(a|e,\theta)$ is uniform