CS674 Natural Language Processing

- Last class
  - MLE
  - Smoothing
    - » Add-one estimation
    - » Witten-Bell discounting
    - » Good-Turing
- Today
  - Combining estimators
    - » Deleted interpolation
    - » Backoff

Combining estimators

- Smoothing methods
  - Provide the same estimate for all unseen (or rare) n-grams
  - Make use only of the raw frequency of an n-gram
- But there is an additional source of knowledge we can draw on --- the n-gram “hierarchy”
  - If there are no examples of a particular trigram, $w_{n-2}w_{n-1}w_n$, to compute $P(w_n|w_{n-2}w_{n-1})$, we can estimate its probability by using the bigram probability $P(w_n|w_{n-1})$.
  - If there are no examples of the bigram to compute $P(w_n|w_{n-1})$, we can use the unigram probability $P(w_n)$.
- For n-gram models, suitably combining various models of different orders is the secret to success.

Simple linear interpolation

- Construct a linear combination of the multiple probability estimates.
  - Weight each contribution so that the result is another probability function.
- Also known as (finite) *mixture models*
- *Deleted interpolation*: when the functions being interpolated all use a subset of the conditioning information of the most discriminating function

$$P(w_n \mid w_{n-2}, w_{n-1}) = \lambda_3 P(w_n \mid w_{n-1}w_{n-2}) + \lambda_2 P(w_n \mid w_{n-1}) + \lambda_1 P(w_n)$$

Backoff (Katz 1987)

- Non-linear method
- The estimate for an n-gram is allowed to back off through progressively shorter histories.
- The most detailed model that can provide sufficiently reliable information about the current context is used.
- Trigram version (first try):

$$\hat{P}(w_i \mid w_{i-2}w_{i-1}) = \begin{cases} P(w_i \mid w_{i-2}w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha_1 P(w_i \mid w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) = 0 \\ & \text{and } C(w_{i-1}w_i) > 0 \\ \alpha_2 P(w_i), & \text{otherwise.} \end{cases}$$

Recursive equation for backoff

$$\hat{P}(w_n \mid w_{n-N+1}^{n-1}) = \tilde{P}(w_n \mid w_{n-N+1}^{n-1}) +$$

$$\theta\,(P(w_n \mid w_{n-N+1}^{n-1}))\,\alpha\,\hat{P}(w_n \mid w_{n-N+2}^{n-1})$$

$$\theta(x) = \begin{cases} 1, & if\ x = 0 \\ 0, & otherwise. \end{cases}$$

$P(.)'s\ are\ MLE$

Backoff + discounting

- Use discounting to tell us how much probability mass to set aside for all the events we haven't see
- Use backoff to tell us how to distribute this probability
- Role of the alphas, P~?
- So…any backoff model must also be discounted/smoothed.

Backoff + discounting

- P-tilda
  - a discounted probability
- Alpha
  - Ensures that the probability mass from all the lower order n-grams sums up to exactly the amount that we saved by discounting the higher-order n-grams

$$\hat{P}(w_n \mid w_{n-N+1}^{n-1}) = \tilde{P}(w_n \mid w_{n-N+1}^{n-1}) +$$

$$\theta\,(P(w_n \mid w_{n-N+1}^{n-1}))\,\alpha(w_{n-N+1}^{n-1})\,\hat{P}(w_n \mid w_{n-N+2}^{n-1})$$

Components

$$\tilde{P}(w_n \mid w_{n-N+1}^{n-1}) = \frac{c^*(w_{n-N+1}^{n})}{c(w_{n-N+1}^{n})}$$

$$\alpha(w_{n-N+1}^{n-1}) = 1 - \sum_{w_n : c(w_{n-N+1}^{n}) > 0} \tilde{P}(w_n \mid w_{n-N+1}^{n-1})$$

normalized by the total probability of all the n-1-grams (bigrams) that begin some n-gram (trigram).

Backoff (final equation)

- Trigram form

$$\hat{P}(w_i \mid w_{i-2}w_{i-1}) = \begin{cases} \tilde{P}(w_i \mid w_{i-2}w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha_1(w_{i-2}^{i-1})\,\tilde{P}(w_i \mid w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) = 0 \\ & \text{and } C(w_{i-1}w_i) > 0 \\ \alpha_2(w_{i-1})\,\tilde{P}(w_i), & \text{otherwise.} \end{cases}$$

Backoff

- When discounting, we usually ignore counts of 1
- Problems?
  - Probability estimates can change suddenly on adding more data when the back-off algorithms selects a different order of n-gram model on which to base the estimate.
- But backoff models are simple, and work well in practice.