Discounting

Jean Mark Gawron

Linguistics
San Diego State University
gawron@mail.sdsu.edu
http://www.rohan.sdsu.edu/∼gawron

Backoff Intuition

1. What happens when an n-gram has 0 counts
2. Back off to (n-1)-gram

Assuming a trigram model

$$(1) \quad P_{\text{katz}}(z \mid x,\ y) = \begin{cases} (a) \quad P^*_{\text{katz}}(z \mid x, y), & \text{if } C(x,\ y,\ z) > 0 \\ (b) \quad \alpha(x, y) P^*_{\text{katz}}(z \mid y) & \text{else if } C(x,\ y) > 0 \\ (c) \quad P^*(z) & \text{otherwise} \end{cases}$$

$$(2) \quad P_{\text{katz}}(z \mid y) = \begin{cases} (a) \quad P^*_{\text{katz}}(z \mid y), & \text{if } C(y,\ z) > 0 \\ (b) \quad \alpha(y) P^*_{\text{katz}}(z) & \text{otherwise} \end{cases}$$

1. Eqn (1) gives the trigram Katz backoff equations, Eqn (2) the bigram Katz backoff eqn. Notice the bigram equation is needed for the RHS of Eqn (1b).
2. Seen events (1a), (2a) must use discounted probabilities P^* because we stealing away probability for the unseen events.
3. α in (1b) and (2b) is a normalization factor explained below.

Katz Backoff

The general formulation of Katz Backoff for any size N:

$$P_{\text{katz}}(w_n \mid w_{n-N+1}^{n-1}) = \begin{cases} P^*_{\text{katz}}(w_n \mid w_{n-N+1}^{n-1}), & \text{if } C(w_{n-N+1}^{n}) > 0 \\ \alpha(w_{i-N+1}^{n-1}) P_{\text{katz}}(w_n \mid w_{n-N+2}^{n-1}) & \text{otherwise} \end{cases}$$

Absolute Discounting

$$\text{P}_{\text{absolute}}(w_{i-1}w_i) = \begin{cases} \dfrac{C(w_{i-1}w_i) - \mathbf{D}}{C(w_{i-1})}, & \text{if } C(w_{i-1}w_i) > 0 \\ \alpha(w_i)P(w_i) & \text{otherwise} \end{cases}$$

1. D is a number (.75) we will subtract from every count.
2. When bigram counts are 0 we use unigram counts, as in Backoff, with a normalization factor α as in Backoff.

Kneser-Ney

Idea: replace the backoff to a unigram probability with a backoff to another kind of probability. Why?

> Unigram models dont distinguish words that are very frequent but only occur in a restricted set of contexts:
>
> San *Francisco*
>
> from words which are less frequent, but occur in many more contexts. The latter may be more likely to finish up an unseen bigram:
>
> I can't see without my reading _____ .
>
> *glasses* is more probable here, but less frequent than Francisco, which almost always occurs after *San*.

The absolute discounting model picks *San*, because it is the word with the higher unigram probability.

Continuation Probability

Kneser-Ney is a little like Witten-Bell in that we pay attention to the number of contexts a word occurs in.

This time, however, it's **preceding** contexts, because we're trying to replace a unigram model with something more informative in backing off.

$$P_{\text{CONTINUATION}}(w_i) == \frac{|\{w_{i-1} : C(w_{i-1}w_i) > 0\}|}{\sum_{w_i} |\{w_{i-1} : C(w_{i-1}w_i) > 0\}|}$$

1. The numerator: the number of word types seen to precede w_i
2. The denominator: the number of word preceding all words.
3. A very frequent word like *Francisco* occuring only in one context (*San*) will have a very low continuation probability.

Kneser-Ney Smoothing

Replace the unigram prob in Absolute discounting with a continuation probability.

$$\mathrm{P}_{\text{absolute}}(w_{i-1}w_i) = \begin{cases} \dfrac{C(w_{i-1}w_i) - \mathbf{D}}{C(w_{i-1})}, & \text{if } C(w_{i-1}w_i) > 0 \\ \alpha(w_i) P_{\textsc{continuation}} & \text{otherwise} \end{cases}$$

Kneser-Ney works even better with something called **interpolation** than it does with backoff:

$$P_{\text{KN}} = \frac{C(w_{i-1}w_i) - D}{C(w_{i-1})} + \beta(w_i) P_{\textsc{continuation}}$$

In interpolation you combine both the discounted bigram prob with a weighted version of the CONTINUATION prob. The weights are functions of w_i and can be set by an HMM training algorithm.