6.864, Fall 2010: Problem Set 1

Total points: 65
Due date: 5 PM, Friday, September 24, 2010
Submit writeup to Tahira Naseem in 32-G362.
Late policy: 5 allowed late days over the course of the semester.

Question 1 (20 points)

In the absolute discounting model of smoothing, all non-zero ML frequencies are discounted by a constant amount δ where $0 < \delta < 1$:
Absolute discounting: If $C(w_n|w_1 \dots w_{n-1}) = r$,

$$P_{abs}(w_n|w_1 \dots w_{n-1}) = \begin{cases} \frac{(r-\delta)}{N} & r > 0 \\ \frac{(V-N_0)\delta}{N_0 * N} & otherwise \end{cases}$$

(Here $C(w_n|w_1 \dots w_{n-1})$ is the number of times $w_1 \dots w_n$ has been seen, P_{abs} is the absolute discounting estimate, V is the size of the vocabulary, N is the total number of times $w_1 \dots w_{n-1}$ has been seen, and N_0 is the number of word types that were unseen after this context.)

Under linear discounting the estimated count of seen words is discounted by a certain fraction, defined by a constant α where $0 < \alpha < 1$.
Linear discounting: If $C(w_n|w_1, \dots, w_n) = r$,

$$P_{lin}(w_n|w_1 \dots w_{n-1}) = \begin{cases} \frac{(1-\alpha)r}{N} & r > 0 \\ \frac{\alpha}{N_0} & otherwise \end{cases}$$

(a) Show that absolute discounting yields a probability distribution for any context $w_1 \dots w_{n-1}$.

(b) Show that linear discounting yields a probability distribution for any context $w_1 \dots w_{n-1}$.

Question 2 (20 points)

Say we have a vocabulary $\mathcal{V}$, i.e., a set of possible words. We'd like to estimate a unigram distribution $P(w)$ over $w \in \mathcal{V}$. We observe n sample points, $w_1, w_2, \dots, w_n$ (this sample may not include all members of $\mathcal{V}$, particularly if n is small compared to $|\mathcal{V}|$.) For any word seen r times in the training sample, the Good-Turing estimate of its count is

$$GT(r) = (r+1) * \frac{N_{r+1}}{N_r},$$

where N_r is the number of members of $\mathcal{V}$ which are seen r times in the corpus. For any w which is observed in the training corpus, we make the estimate $P(w) = GT(C(w))/n$, where $C(w)$ is the number of times w is seen in the sample.

(a) Can you see any problem with this estimation method for words with large values for $C(w)$?

(b) Prove that under this definition $\sum_{w \in \mathcal{V}'} P(w) \leq 1$, where $\mathcal{V}'$ is the subset of $\mathcal{V}$ seen in the training corpus. If the "missing" probability mass $1 - \sum_{w \in \mathcal{V}'} P(w)$ is divided evenly amongst the words not seen in the corpus, show that $P(w)$ for any word not in the corpus is $N_1/(n \times N_0)$ where N_0 is $|\mathcal{V}| - |\mathcal{V}'|$, and N_1 as before is the number of members of $\mathcal{V}$ seen exactly once in the corpus. (You can assume that $N_r > 0$ for $r = 1, \ldots, k$ for some $k > 1$ and $N_r = 0$ for $r > k$.)

Question 3 (25 points)

We train a trigram language model using add-α smoothing on a large corpus composed of Wall Street Journal (WSJ) articles from 2003.

(a) Plot the shape of the probability of your **training** corpus under the resulting language model as a function of α for $0 \leq \alpha < \infty$.

(b) We now test this language model on WSJ articles from 2004. Plot the probability of your **test** corpus under this language model as a function of α.

(c) Plot the *perplexity* on the test corpus under this language model as a function of α.

(d) Let V be the size of the vocabulary and W be the size of the test corpus, both in number of words. As $\alpha \to \infty$, what value does the perplexity on the test set approach? Explain your answer.