Information Extraction from Speech and Text

Homework # 7

Due April 7, 2011.

Review Chapter 15 from Statistical Methods for Speech Recognition by Frederick Jelinek.

1. **Comparison of Held-Out and Good-Turing Estimates**: We will compare the held-out estimate of Section 15.2 and the Good-Turing estimate of Section 15.4 by using text A and text B from our projects. In particular, we will use text A to develop our probability estimates, and text B to check their empirical performance in terms of the average log-likelihood or perplexity. For this task, the alphabet $X$ will be non-overlapping pairs of consecutive letters, so that $|X| = 729$, $|A| = 15,000$ and $|B| = 2,500$. We will assume that consecutive symbols are independent, and estimate unigram probability distributions on $X$.

   (a) Let the first 12,000 symbols in text A be the development set $D$, and the remaining 3,000 the held-out set $H$. Use the procedure described in Section 15.2 to

   i. begin with a provisional value of $M = 20$;
   ii. use the counts $c_d(x)$ in $D$ to create equivalence classes $\Phi : X \rightarrow \{0, 1, \ldots, M+1\}$, i.e $\Phi(x) = i$ if and only if $c_d(x) = i$;
   iii. use the counts $r_i$ in $H$ to estimate the class-probabilities $\lambda_i$, $i = 0, 1, \ldots, M + 1$;
   iv. use the counts $c_d(x)$ in $D$ to get the class membership counts $n_i$, $i = 0, 1, \ldots, M$, the relative frequency estimates $f_d(x)$ for $x$ with $\Phi(x) = M+1$, and the probability $P_M$;
   v. use the considerations in Section 15.2.3 to choose an appropriate value of $M$ in Step 1(a)i above, then repeat Steps 1(a)ii through 1(a)iv;
   vi. compute the probability estimate

   $\tilde{P}(x) = \begin{cases} 
   \lambda_i \times \frac{1}{n_i} & \text{if } \Phi(x) = i \in \{0, 1, \ldots, M\}, \\
   \lambda_{M+1} \times \frac{f_d(x)}{P_M} & \text{if } \Phi(x) = M + 1,
   \end{cases} \quad x \in X, \quad (1)$

   and verify that it sums to unity;
   vii. compute the perplexity of text B using the held-out estimate of equation (1);
   viii. *(alternative)* just before Step 1(a)vi, pool the data-sets $D$ and $H$ back together and use the counts $c_d(x)$ in $D \cup H$ to recalculate the equivalence classes $\Phi : X \rightarrow \{0, 1, \ldots, M + 1\}$, the class membership counts $n_i$, the estimates $f_d(x)$ and the probability $P_M$ in (1).
Compare the performance of the estimate with and without the optional Step 1(a)viii of merging together \( D \) and \( H \) for reestimating the class membership and relative frequencies.

(b) Good-Turing estimation does not require dividing the text \( A \) into \( D \) and \( H \). Therefore,

i. begin with a provisional value of \( M \) as determined in Step 1(a)v above;

ii. compute the counts \( c_d(x) \), equivalence classes \( \Phi : \mathcal{X} \rightarrow \{0, 1, \ldots, M + 1\} \), the class membership counts \( n_i \), the estimates \( f_d(x) \) and \( P_M \) from the entire text \( A \);

iii. using \( N \) to denote \( |A| \), compute the probability estimate

\[
\hat{P}(x) = \begin{cases} 
\frac{(i+1)n_{i+1}}{n_i N} & \text{if } \Phi(x) = i \in \{0, 1, \ldots, M\}, \\
\alpha f_d(x) & \text{if } \Phi(x) = M + 1,
\end{cases} 
\quad x \in \mathcal{X},
\tag{2}
\]

where \( \alpha \) is computed, as described in Section 15.4, to ensure that \( \hat{P}(\cdot) \) sums to unity.

iv. compute the perplexity of text \( B \) using the Good-Turing estimate of equation (2).

How should one choose \( M \) in Step 1(b)i above? Experiment with a few different values.

Compare the perplexity of the best Good-Turing estimate of Part 1b with that of the best held-out estimate of Part 1a.

2. Comparison of Kneser-Ney and Katz Back-off: Read the article


Contrast the lower order distributions proposed in this article with the Katz backoff model in Section 15.7, and interpret the differences in your own words. Explain via illustrative examples whether you expect one to be a better model for naturally occurring text than the other.