Natural Language Processing

Problem Set 1: Analytical

Linan Qiu (lq2137)

September 22, 2014

1 Linear Interpolation

Expanding the linear interpolation parameter function,

$$
\begin{aligned}
L(\lambda_1, \lambda_2, \lambda_3) &= \sum_{w_1, w_2, w_3} c'(w_1, w_2, w_3) \log q(w_3|w_1, w_2) \\
&= \sum_{w_1, w_2, w_3} c'(w_1, w_2, w_3) \log \left(\lambda_1 q_{ML}(w_3|w_1, w_2) + \lambda_2 q_{ML}(w_3|w_2) + \lambda_3 q_{ML}(w_3)\right)
\end{aligned}
$$

On the other hand, perplexity is defined as 2^{-l} where

$$l = \frac{1}{M} \sum_{i=1}^{m} \log p(x^i)$$

where m is the number of sentences and $p(x^i)$ is the probability of sentence x^i occurring.

Minimizing perplexity means maximizing l. Expanding l for trigram models, where $p(x^i_j)$ is the probability of word x^i_j occurring in the sentence i

$$
\begin{aligned}
l &= \frac{1}{M}\sum_{i=1}^{m}\log p(x^i) \\
&= \frac{1}{M}\sum_{i=1}^{m}\Big(\log p(x_1^i) + \log p(x_2^i) + ... + \log p(x_n^i)\Big) \\
&= \frac{1}{M}\sum_{i=1}^{m}\sum_{j=1}^{n}\log p(x_j^i) \\
&= \frac{1}{M}\sum_{i=1}^{m}\sum_{j=1}^{n}\log\Big(\lambda_1 q_{ML}(x_j^i|x_{j-2}^i, x_{j-1}^i) + \lambda_2 q_{ML}(x_j^i|x_{j-1}^i) + \lambda_3 q_{ML}(x_j^i)\Big)
\end{aligned}
$$

To minimize perplexity would mean to maximize this with respect to $\lambda_1, \lambda_2, \lambda_3$.

$$
\max_{\lambda_1,\lambda_2,\lambda_3}\frac{1}{M}\sum_{i=1}^{m}\sum_{j=1}^{n}\log\Big(\lambda_1 q_{ML}(x_j^i|x_{j-2}^i, x_{j-1}^i) + \lambda_2 q_{ML}(x_j^i|x_{j-1}^i) + \lambda_3 q_{ML}(x_j^i)\Big)
$$

Since this includes duplicates, we can shorten portion to the right of $\sum_{j=1}^{n}$ to

$$
\begin{aligned}
&\max_{\lambda_1,\lambda_2,\lambda_3}\frac{1}{M}\sum_{i=1}^{m}\sum_{x_{j-2}^i, x_{j-1}^i, x_j^i} c'(x_{j-2}^i, x_{j-1}^i, x_j^i)\log\Big(\lambda_1 q_{ML}(x_j^i|x_{j-2}^i, x_{j-1}^i) + \lambda_2 q_{ML}(x_j^i|x_{j-1}^i) + \lambda_3 q_{ML}(x_j^i)\Big) \\
&= \max_{\lambda_1,\lambda_2,\lambda_3} L(\lambda_1, \lambda_2, \lambda_3)
\end{aligned}
$$

Hence, maximizing the parameters means maximizing l which means minimizing the perplexity.

2 Linear Interpolation with Bucketing

Defining Φ as a mapping of **trigram** into bins is erroneous as it will produce λs that do not sum up to 1.

For a text sequence y_{i-2}, y_{i-1}, y_i, where

$$
\begin{aligned}
Count(y_{i-2}, y_{i-1}, y_i) &= 0 \\
Count(y_{i-1}, y_i) &= 0
\end{aligned}
$$

We have to ensure that $\lambda_1^{\Phi(y_{i-2},y_{i-1},y_i)} = 0$ since $Count(y_{i-1}, y_i) = 0$ and the trigram will be undefined.

However, the following counts

$$
\begin{aligned}
Count(y_{i-2}, y_{i-1}, y_i) &= 0 \\
Count(y_{i-1}, y_i) &> 0
\end{aligned}
$$

will also generate the same $\lambda_1^{\Phi(y_{i-2},y_{i-1},y_i)} = 0$. However, we would not want $\lambda_1^{\Phi(y_{i-2},y_{i-1},y_i)} = 0$ since the trigram is still defined and $\lambda_1^{\Phi(y_{i-2},y_{i-1},y_i)}$ should have a non-zero weight.

Hence there is no way to ensure that the λs will sum up to 1. We violate the constraint that $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

3 Modified Viterbi

- **Input:** a sentence $x_1 ... x_n$, parameters $q(s|u, v)$ and $e(x|y)$
- **Definitions:** Define $T(x)$ to be the tag dictionary that lists the tags y such that $e(x|y) > 0$. Define S to be the set of possible tags. Define $S_{-1} = S_0 = \{*\}$. Define $S_k = T(x_k)$ for $k = 1...n$
- **Initialization:** Set $\pi(0, *, *) = 1$
- **Algorithm:**
  - For $k = 1...n$
    - For $u \in S_{k-1}$, $v \in S_k$,

$$
\begin{aligned}
\pi(k, u, v) &= \max_{w \in S_{k-2}} (\pi(k-1, w, u) * q(v|w, u) * e(x_k|v)) \\
bp(k, u, v) &= \arg\max_{w \in S_{k-2}} (\pi(k-1, w, u) * q(v|w, u) * e(x_k|v))
\end{aligned}
$$

  - Set $(y_{n-1}, y_n) = \arg\max_{u \in S_{n-1}, v \in S_n} (\pi(n, u, v) * q(STOP|u, v))$
  - For $k = (n-2)...1$,

$$
y_k = bp(k+2, y_{k+1}, y_{k+2})
$$

- **Return** the tag sequence $y_1 ... y_n$