COMS 4705: Natural Language Processing HW1

Zhuangyu Ren(zr2209)

February 12, 2019

Question 1

perplexity $= 2^{-l}$, where

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

As $p(s_i) = \prod_{i=1}^{n} q(x_i|x_{i-2}, x_{i-1})$, where n is the number of words in a sentemce
We now have

$$\begin{aligned} l &= \frac{1}{M} \sum_{i=1}^{m} \log \prod_{j=1}^{n} q(x_j|x_{j-2}, x_{j-1}) \\ &= \frac{1}{M} \sum_{i=1}^{m} \sum_{j=1}^{n} \text{log} q(x_j|x_{j-2}, x_{j-1}) \end{aligned}$$

Here, m is the total number of sentences and n is the number of words in each sentence. We can distract all the trigrams in all sentences, and the total number of all pairs of trigrams adds up to M. If we combine all the sentences and use $c'(w_1, w_2, w_3)$ to describe the number of that particular trigram, we will have

$$l = \frac{1}{M} \sum_{w_1, w_2, w_3} c'(w_1, w_2, w_3) \text{log} q(x_j|x_{j-2}, x_{j-1})$$

We want to minimize the perplexity, which means that we need to maximize l. And the formula equation above can be represented as

$$l = \frac{1}{M} \sum_{w_1, w_2, w_3} c'(w_1, w_2, w_3) \text{log} q(x_j|x_{j-2}, x_{j-1}) = \frac{1}{M} L(\lambda_1, \lambda_2, \lambda_3)$$

So choosing λ values that maximize $L(\lambda_1, \lambda_2, \lambda_3)$ is equivalent to choosing λ values that minimize the perplexity of the language model on the validation data.

Question 2

For all w_i given w_{i-2}, w_{i-1}, we have $\sum_{w_i} q(w_i|w_{i-2}, w_{i-1}) = 1$, we will then prove this.

$$\begin{aligned}\sum_{w_i} q(w_i|w_{i-2}, w_{i-1}) &= \sum_{w_i}(\lambda_1^{\Phi(w_{i-2},w_{i-1})} q_{ML}(w_i|w_{i-2}, w_{i-1}) + \lambda_2^{\Phi(w_{i-2},w_{i-1})} q_{ML}(w_i|w_{i-1}) \\ &+ \lambda_3^{\Phi(w_{i-2},w_{i-1})} q_{ML}(w_i)) \\ &= \sum_{w_i} \lambda_1^{\Phi(w_{i-2},w_{i-1})} q_{ML}(w_i|w_{i-2}, w_{i-1}) + \sum_{w_i} \lambda_2^{\Phi(w_{i-2},w_{i-1})} q_{ML}(w_i|w_{i-1}) \\ &+ \sum_{w_i} \lambda_3^{\Phi(w_{i-2},w_{i-1})} q_{ML}(w_i)\end{aligned}$$

Because $\Phi(w_{i-2}, w_{i-1})$ is independent with w_i, we can take out the λ, so

$$\sum_{w_i} q(w_i|w_{i-2}, w_{i-1}) = \lambda_1^{\Phi(w_{i-2},w_{i-1})} \sum_{w_i} q_{ML}(w_i|w_{i-2}, w_{i-1}) + \lambda_2^{\Phi(w_{i-2},w_{i-1})} \sum_{w_i} q_{ML}(w_i|w_{i-1}) \\ + \lambda_3^{\Phi(w_{i-2},w_{i-1})} \sum_{w_i} q_{ML}(w_i)$$

Take $q_{ML}(w_i|w_{i-2}, w_{i-1})$ as an example:

$$q_{ML}(w_i|w_{i-2}, w_{i-1}) = \frac{count(w_{i-2}, w_{i-1}, w_i)}{count(w_{i-2}, w_{i-1})}$$

So, given $w_{i-2}, w_{i-1}, \sum_{w_i} q_{ML}(w_i|w_{i-2}, w_{i-1}) = 1$.
Same for $q_{ML}(w_i|w_{i-1})$ and $q_{ML}(w_i)$, they all sum to 1.
Now,

$$\sum_{w_i} q(w_i|w_{i-2}, w_{i-1}) = \lambda_1^{\Phi(w_{i-2},w_{i-1})} + \lambda_2^{\Phi(w_{i-2},w_{i-1})} + \lambda_3^{\Phi(w_{i-2},w_{i-1})} = 1$$

But now we have a function Φ, which is relavent to w_i, so we cannot take λ out of the whole equation. So we can not have the new expression

$$q(w_i|w_{i-2}, w_{i-1} = \lambda_1^{\Phi(w_{i-2},w_{i-1},w_i)} q_{ML}(w_i|w_{i-2}, w_{i-1}, w_i) + \lambda_2^{\Phi(w_{i-2},w_{i-1})} q_{ML}(w_i|w_{i-1}) + \lambda_3^{\Phi(w_{i-2},w_{i-1},w_i)} q_{ML}(w_i)$$

sum up to one.

Question 3

Input: a word sequence $x_1...x_n$, a tag dictionary T(x) that lists the tags y such that $e(x|y) > 0$

Initialization: Set $\pi(0, *, *) = 1$;
Definition: $S_{-1} = S_0 = \{*\}, S_k = T$ for $k \in \{1...n\}$;
for $k = 1...n$ **do**
    for $u \in S_{k-1}, v \in S_k$ **do**
        $\pi(k, u, v) = \underset{w \in S_{k-2}}{max}(\pi(k-1, w, u) \times q(v|w, u) \times e(x_k|v))$;
    end
end
Return: $\underset{u \in S_{n-1}, v \in S_n}{max}(\pi(n, u, v) \times q(STOP|u, v))$;

Algorithm 1: Viterbi algorithm