HW 1.1 report

b01504044 化工五 宋易霖

問題簡述:

題目給定一段文字，這串文字是用一個有意義的英文文章(通常)，經過一個 encode function 所轉換而來的。而我們需要利用題目給定的 probability encode function 和 bigram table 來找出原本的文章文字。

註:bigram 的意思是文章中連續兩個字母出現的頻率，例如: 本題中 gg 出現的頻率是 0.22、gt 出現的頻率是 0.004。

問題假設

根據提議，我的 graphic model 有以下基本假設:

1. 每個字元只跟前一個字元和 encode 後的字元相關
2. 每個單字結尾都有空白(包括最後一個單字)

解題方法:

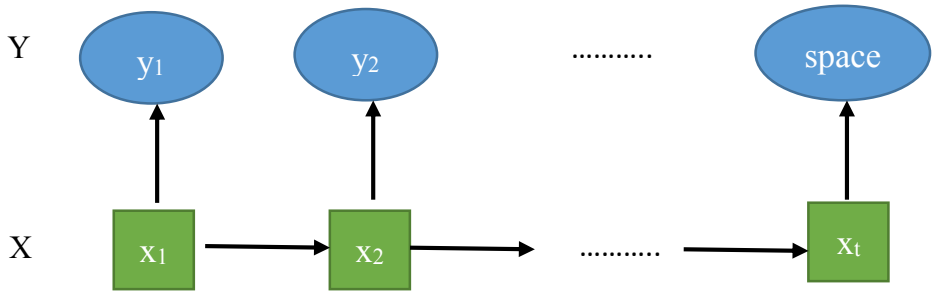
1. Graphic model and inference:

因為以上假設具有方向性，因此 graphic model 為一 bayesian model，每個字元為一個 node，每個字元 x_t 都和 x_{t-1}、y_t 連接，edge 方向為 x_{t-1} 指向 x_t，x_t 指向 y_t。

而我的 model 是一個單字一個單字看的，因為用一整篇文章一起看跟分單字看結果會一樣: 當遇到 y 是空白的時候，不管前面是甚麼字元，這格的 x 必定是空白，而空白的下一個字元的前一個字元就會都是空白，這是除了第一個單字以外都會有的共同特徵。又因為字元只跟前一個字元有關(先不考慮 y)，所以每個單字的第一個字元只會跟前面的空白有關係，不會跟前一個單字的結尾有關(也就是兩個單字彼此獨立)，因此可以將單字拆分開來看而不會影響準確度。而且分單字看還能節省記憶體，因此選擇分單字處理。

但是仍會因為全文的單字位置而有不同處理:

a. 全文章中第一個字(下圖每一個框框表示一個字元)



HW 1.1 report

b01504044 化工五 宋易霖

Inference:

$$MPA = \max_{x_1, x_2 \ldots x_t} P(x_1, x_2 \ldots x_t y_1, y_2 \ldots \ldots y_t = \text{space})$$

$$= \max_{x_t} P(y_t = \text{space} \mid x_t) P(x_t \mid x_{t-1}) \max_{x_{t-1}} P(y_{t-1} \mid x_{t-1}) P(x_{t-1} \mid x_{t-2}) \ldots\ldots \max_{x_1} P(y_1 \mid x_1) P(x_1)$$

公式解析:

I. 對第一個字元來說，因為是全文章第一個字，前方沒有空白，所以此時 x 的機率全由 $P(x_1, y = y_1)$決定:

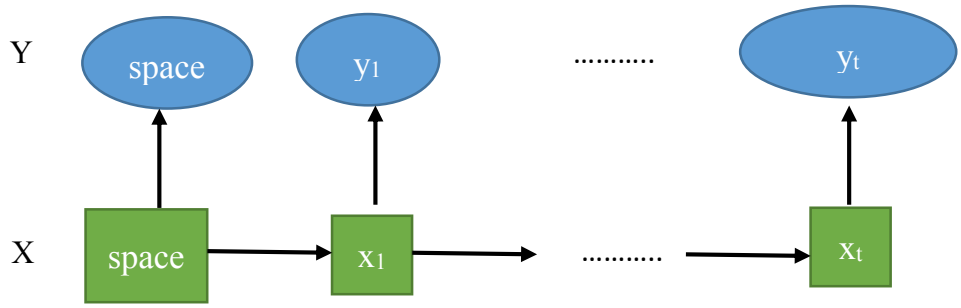
$\max_{x_1} P(y_1 \mid x_1) P(x_1) = \max_{x_1} P(x_1, y_1)$ 即每個 x_1 的機率為對應到 encode 中 x_1, y_1 那格中的機率。

II. 對後面的字元來說，x_t 的機率被前一個字 x_{t-1} 和 y_t 共同決定:

$$\max_{x_t} P(y_t \mid x_t) P(x_t \mid x_{t-1}) P(x_{t-1}) = \max_{x_t} P(y_t \mid x_t) P(x_{t-1}, x_t)$$

用上述公式求出 x_t 最大的機率和將使 x_t 成為最大機率的 x_{t-1} 存起來。
而這個步驟需要做到單字後面的空白，因為每個字元變成空白的機率不同，因此仍舊要考慮這一步。

b. 文章中最後一字



Inference:

$$MPA = \max_{x_0, x_1 \ldots x_t} P(x_0, x_1, x_2 \ldots . x_t, y_0 = \text{space}, y_1 \ldots \ldots y_t)$$

$$= \max_{x_t} P(y_t \mid x_t) P(x_t \mid x_{t-1}) \max_{x_{t-1}} P(y_{t-1} \mid x_{t-1}) P(x_{t-1} \mid x_{t-2}) \ldots\ldots \max_{x_0} P(y_0 = \text{space} \mid x_0) P(x_0)$$

公式解析:

I. 對一個字元來說，前方有空白，所以此時 x 的機率由

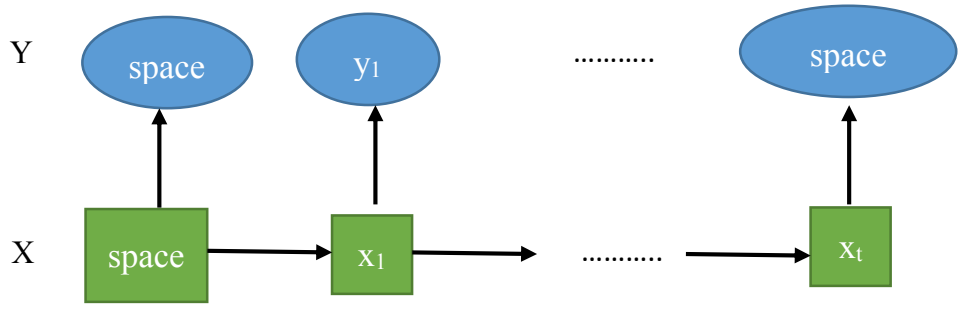
$\max_{x_1} P(y_1 \mid x_1) P(x_1 \mid x_0) P(x_0) = \max_{x_1} P(y_1 \mid x_1) P(x_0 = \text{space}, x_1)$決定。

HW 1.1 report

b01504044 化工五 宋易霖

II. 其他字元情況和 a.中的 II 類似，只不過最後沒有空白，所以做到該單字的最後一個字元就可以結束了。

c. 文章中其他字



Inference:

$$MPA = \max_{x_0, x_1 \ldots x_t} P(x_0, x_1, x_2 \ldots x_t, y_0 = \text{space}, y_1 \ldots\ldots y_t = \text{space})$$

$$= \max_{x_t} P(y_t = \text{space} | x_t) P(x_t | x_{t-1}) \max_{x_{t-1}} P(y_{t-1} | x_{t-1}) P(x_{t-1} | x_{t-2}) \ldots\ldots \max_{x_0} P(y_0 = \text{space} | x_0) P(x_0)$$

公式解析:

I. 和 b.I 的狀況處理一樣。

II. 其他字元情況和 a.中的 II 一樣。

2. 輸出單字

把每個單字的所有字元的機率都做好之後，選擇最後一個字元中的機率最大的那個開始回溯，並且輸出。因為在處理時都有存下前一個字元，所以就像 dynamic programming 一樣的做法即可。

參考資料:

1. 上課講義 L3_Inference.pptx p.51~P.68

2. 和 R05922027 江東峻 同學討論