

基於隱馬可夫模型之中文新聞語意結構分析與分類系統

A Chinese News Semantic Analysis and Classification System based on Hidden Markov Model

指導教授：賀保羅
組員：賴志宇

Introduction

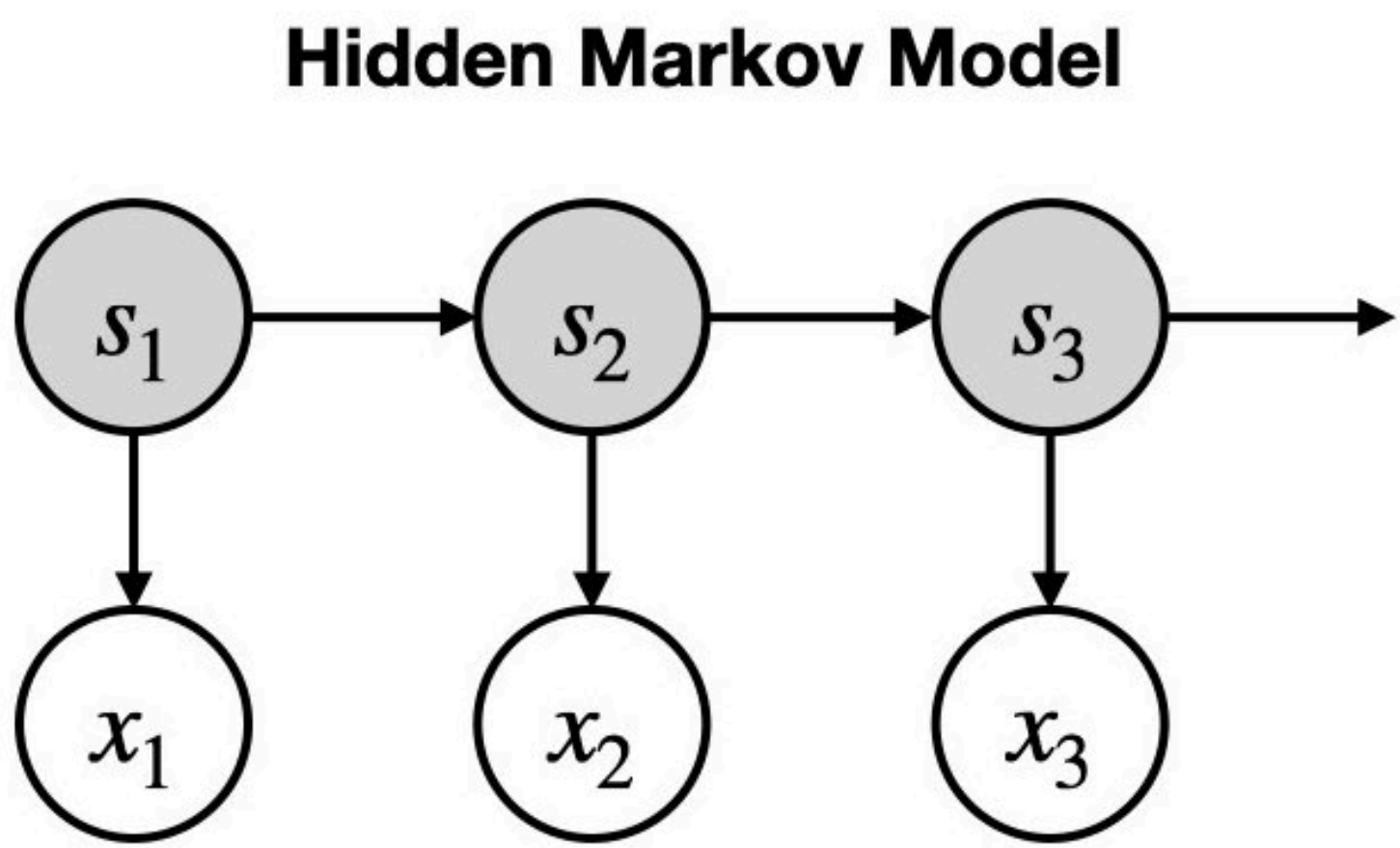
隱馬可夫模型HMM:

HMM認為觀測序列背後有一個不可見的狀態序列，狀態序列在下一個時刻會根據狀態轉移矩陣轉移其狀態，以及根據發射機率矩陣發射出我們實際看到的觀測值。

系統概念：

中文文章背後亦有語意結構等等，將語意結構的概念映射到隱狀態；觀察到的字詞映射到序列，我們就可以透過隱馬可夫模型打造一個分析系統。

相對於大部分基於深度學習的NLP系統而言，使用HMM的好處在於系統的可解釋性強，且能夠看出不同種類文章的語意結構、轉折等等。



Algorithm

$$\lambda = (A, B, \pi)$$
$$A = [a_{ij}], \quad a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$$
$$B = [b_j(k)], \quad b_j(k) = P(v_k \mid q_t = S_j)$$
$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \cdot b_j(o_t)$$

Architecture

數據處理：
抓14類資料、分詞
將字詞轉化為id序列，出現次數<20設為null

模型訓練：
針對14類新聞分別進行
隱狀態數目為8/12/16/20
Baum-Welch訓練(n_iter = 50)

使用者介面：
讀取文本
計算各類模型對文本的分數
用Viterbi演算法還原路徑

Model Accuracy

Category	Test count	Correct count	accuracy
时尚	200	198	99.00%
家居	200	184	92.00%
教育	200	195	97.50%
股票	200	191	95.50%
娱乐	200	199	99.50%
彩票	200	188	94.00%
社会	200	186	93.00%
房产	200	186	93.00%
星座	200	199	99.50%
科技	200	187	93.50%
财经	200	186	93.00%
时政	200	194	97.00%
游戏	200	194	97.00%
体育	200	199	99.50%

Overall Accuracy: 95.93%

Demo

HMM 新聞結構分析系統

美軍連兩日空襲伊朗，川普兩週未獲回覆轉趨強硬
【財訊快報 / 陳孟翔】美伊軍事對峙週三再度升級！美國中央司令部宣布，美東時間週三下午5時15分(台北時間週四清晨5時15分)，依據美國總統川普指示，對伊朗境內多個目標發動追加「自衛性」打擊，以回應伊朗持續挑釁。這是美軍連續第二天對伊朗發動攻擊，目標集中在伊朗南部，包括防空系統、雷達，以及無人機指揮與控制單元。消息人士指出，川普在等待近兩週後，仍未收到伊朗對美方最新協議提案的回應，對談判拖延愈來愈沮喪。川普在白宮表示，美國已準備好一份可確保伊朗永遠無法擁有核武的新協議，但伊朗遲遲不願簽署。他並稱，美方此次出動B-2隱形轟炸機對伊朗進行打擊，意在迫使德黑蘭重新評估談判立場，而非放任衝突無限升級。

開始分析

預測類別：時政

