D10 詞性標註(Part-of-speech tagging)

簡報閱讀 範例與作業 問題討論

重要知識點

- 了解詞性標注原理以及應用

資料來源：http://parts-of-speech.info/

Part Of Speech (POS)

- 詞類(POS)：每種語言都由許多詞類組成，例如動詞，名詞，副詞，形容詞等。
- 詞性標註(Part Of SpeechTagging)，簡單來說就是將文章、句子中，文字的詞類標註出來，為 NLP 任務中相當重要的技術之一。

資料來源：Categorizing and POS Tagging with NLTK Python

標註原理

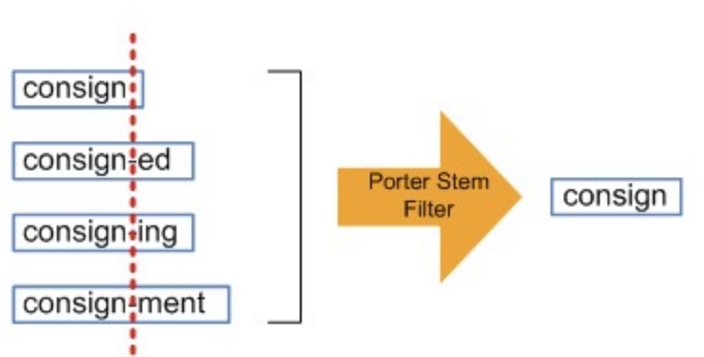
- POS 標註任務中，信息擴展基於詞本身的**內在信息**和基於某些的**外在信息**，換而言之，當我們在決定單詞的詞性前，除了考慮單詞本身，也要考慮前後單詞與整句話。
- 通常一個單詞會包含多種詞性。

標註意義

- 詞性標註能在許多 NLP 的任務中提供低層次的語義信息。

POS-tagging 應用

- 用於模型輸入特徵：提供單詞與鄰近單詞的訊息，以利進一步分析與處理。
- 提供句法結構的訊息，可用來做相似度判斷等應用。
- 詞幹提取(stemming)：去除詞綴得到詞根

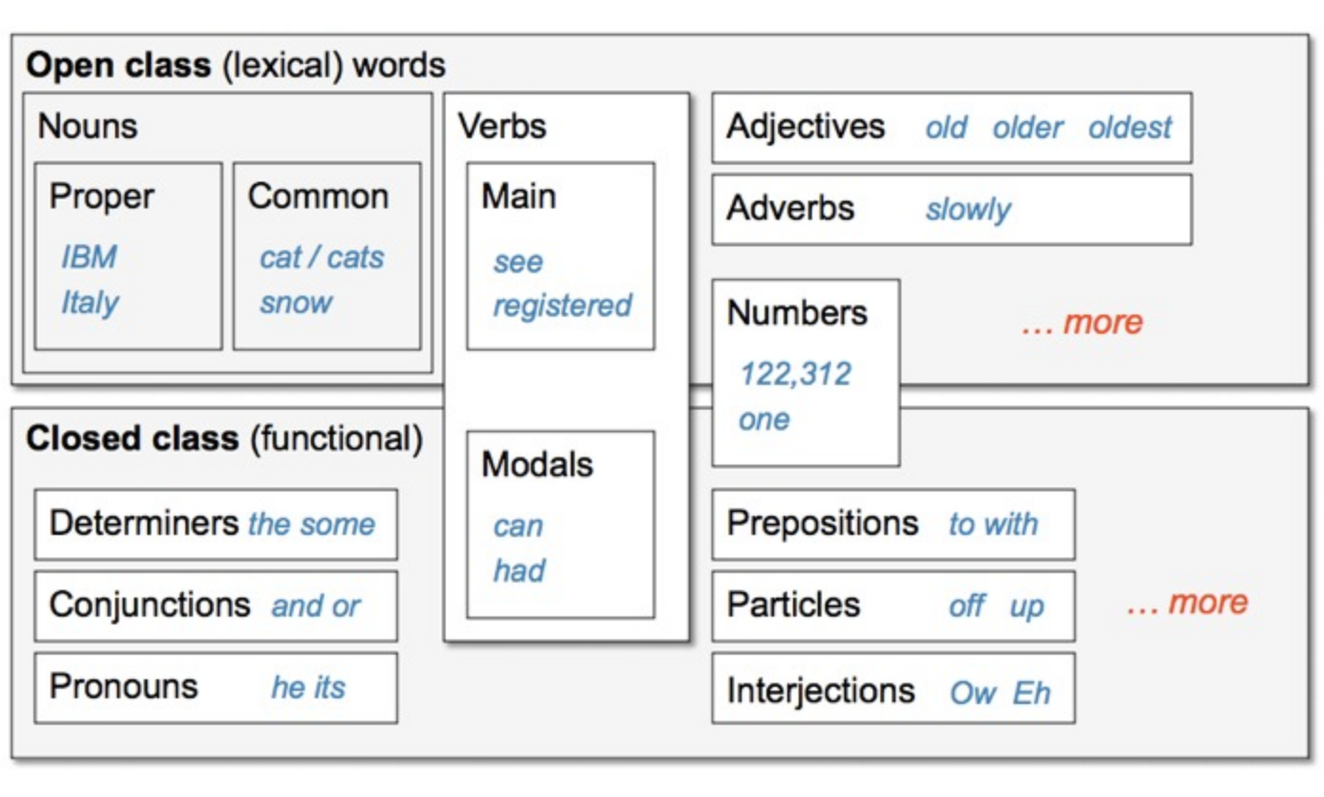


資料來源：去字尾的方法，使用Porter Stemming Algorithm

POS-分類

主要可分為兩類：open class，closed class

- closed class：通常為相對固定的詞類，不太會有新的詞類出現
  Ex. pronouns: she, he, I
  preposition: on, under, by
- open class：容易有新詞被創造，如名詞、動詞、形容詞等等



資料來源：12 1 An Intro to Parts of Speech and POS Tagging

POS Tagging 常見算法

- Lexical Based Methods：直接使用訓練詞庫中該單詞最常見的詞性作為標註。
- Rule-Based Methods ：使用自訂的 rules 來標記單詞，如看到 ed、i 就標注 verb。
- Probabilistic Methods ：使用條件機率的原理，預測單詞詞性，常見如 CRF、HMM，此方法也是深度學習出來前，最常見且效果最好的標注方式。
- Deep Learning Methods ：使用深度學習模型預測標註詞性。

資料來源：12 1 An Intro to Parts of Speech and POS Tagging

> **在Pos Tagging中，Probabilistic Methods 是最常見且效果相當好的一種方式，其中又以HMM最為常見。**

HMM(隱藏式馬可夫模型 Hidden Markov Model)

HMM 主要可以用來解決三種經典的問題

- **預測(filter)**：已知模型參數和某一特定輸出序列，求最後時刻各個隱含狀態的機率分布，通常使用**前向演算法**解決
- **平滑(smoothing)**：已知模型參數和某一特定輸出序列，求中間時刻各個隱含狀態的機率分布，通常使用**前向-後向演算法**解決
- **解碼(most likely explanation)**：已知模型參數，尋找最可能的能產生某一特定輸出序列的隱含狀態的序列，通常使用**Viterbi演算法**解決

資料來源：隱藏式馬可夫模型

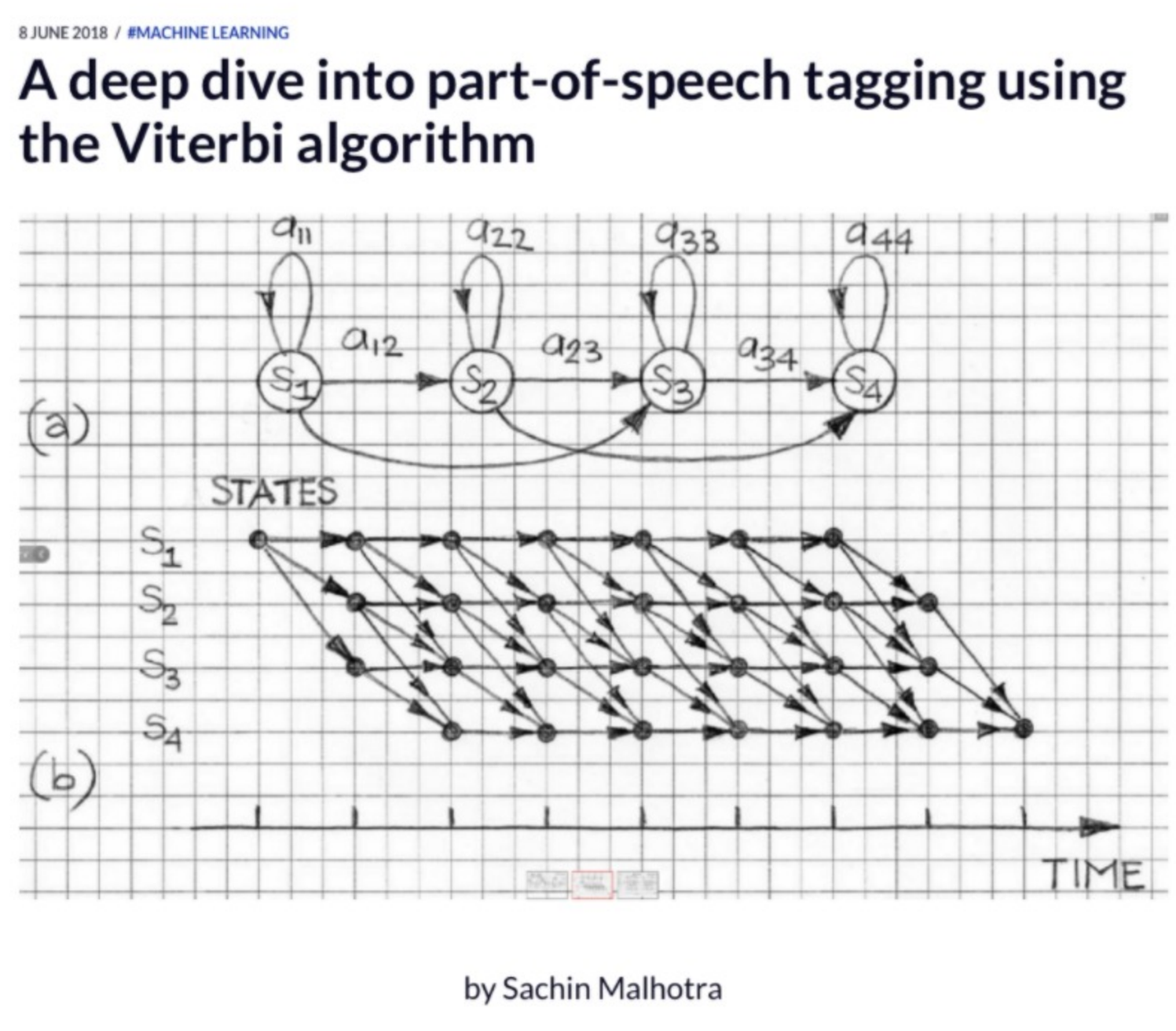
HMM for PoS Tagging

將 HMM 應用在 Pos Tagging 中，符合解碼的問題

- **解碼(most likely explanation)**：已知模型參數，尋找最可能的能產生某一特定輸出序列的隱含狀態的序列，通常使用**Viterbi演算法**解決

> **也就是當我們知道完整句子時，我們如何推論出最有可能的 Tagging 序列。**

想要詳細理解 HMM 算法的學員可以參考這篇文章：
網站：A deep dive into part-of-speech tagging using the Viterbi algorithm

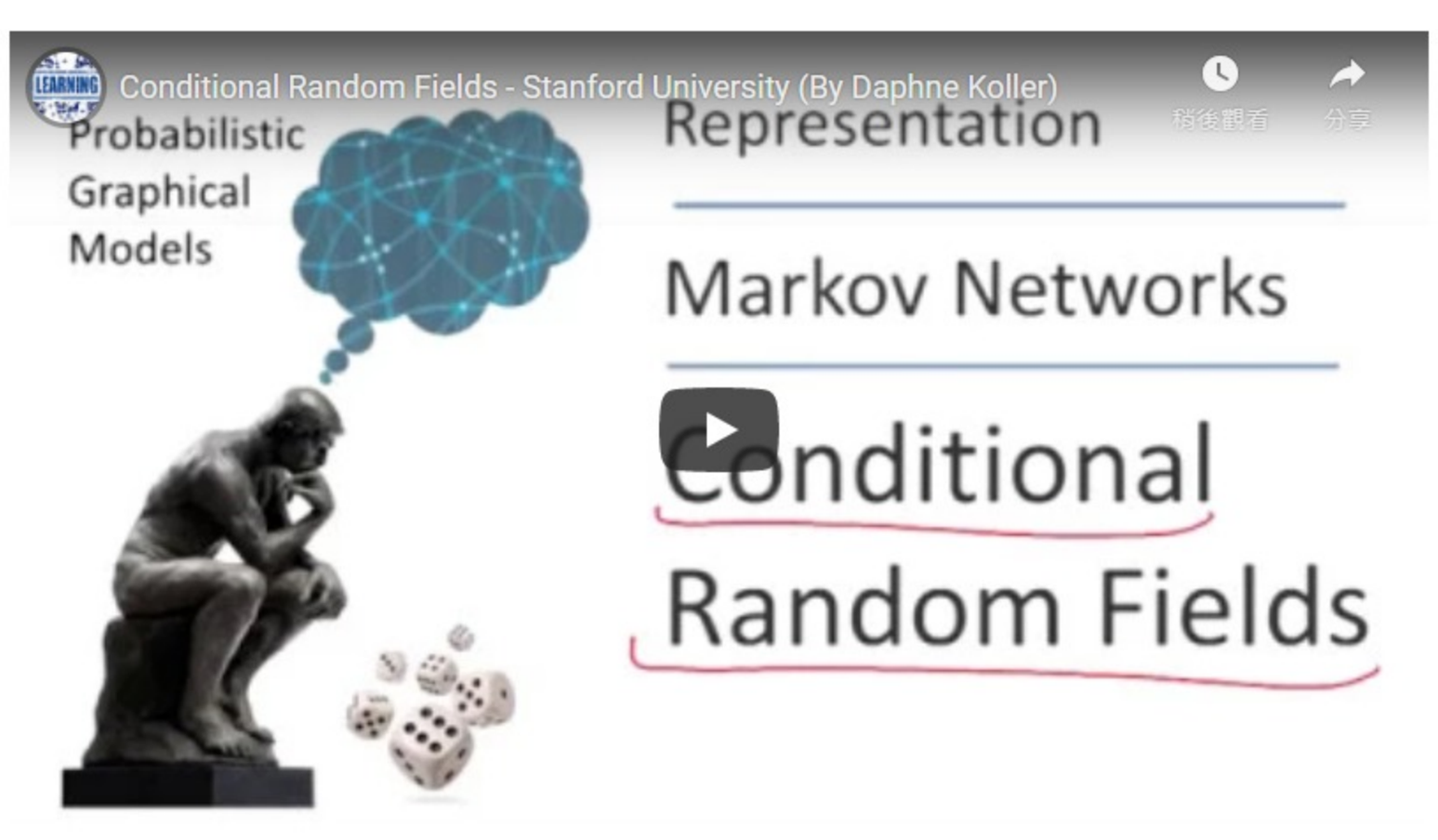


by Sachin Malhotra and Divya Godayal

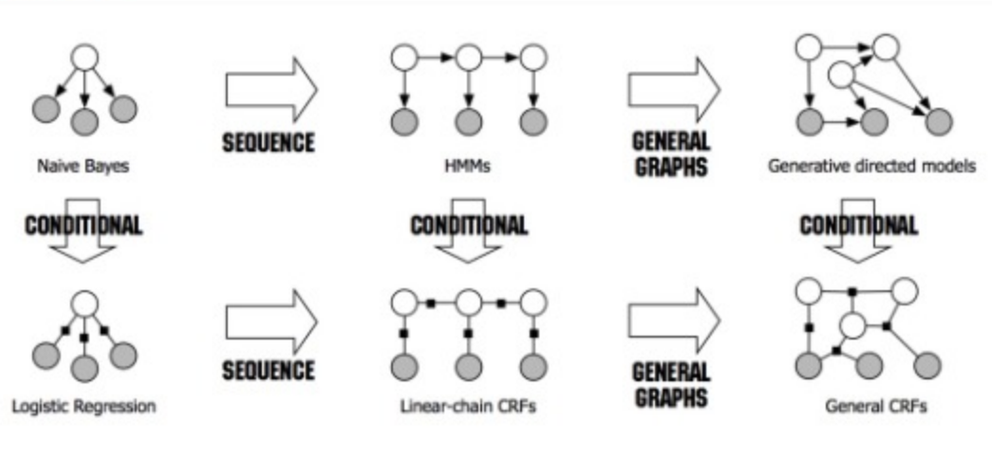
參考資料

網站：Christopher Manning : POS-tagging
Christopher Manning 講解 POS 的影片

網站：CRF原理解釋影片：Daphne Koller



網站：CRF解釋：中文



條件隨機場(Conditional Random Fields, CRFs)筆記

谢少

2 人赞同了该文章

最近在proposal的時候，口試委員提出可以用CFRs模型。起初，只是聽過，在statistical learning中看過，也沒去了解（畢竟不會用得到）。但是，口試委員提起了，那就要看。以防再次被問及，答不出來。這兩天花時間看了一下，簡要的做一下筆記。

上面的封面來自常被提及的《An Introduction to Conditional Random Fields》By Charles Sutton and Andrew McCallum。我不會全部提及圖中的所有內容，畢竟知乎上有很多相關的內容。

https://homepages.inf.ed.ac.uk/csutton/publications/crftut-fnt.pdf
homepages.inf.ed.ac.uk

一、隱形馬爾科夫模型（Hidden Markov Models，HMMs）

HMMs是建立狀態序列和觀測序列的一個生成模型(Generative Model)。因為是生成模型，所以它是建立在聯合概率下的模型，即：p(觀測序列，狀態序列)。

下一步：閱讀範例與完成作業