藏頭詩產生器

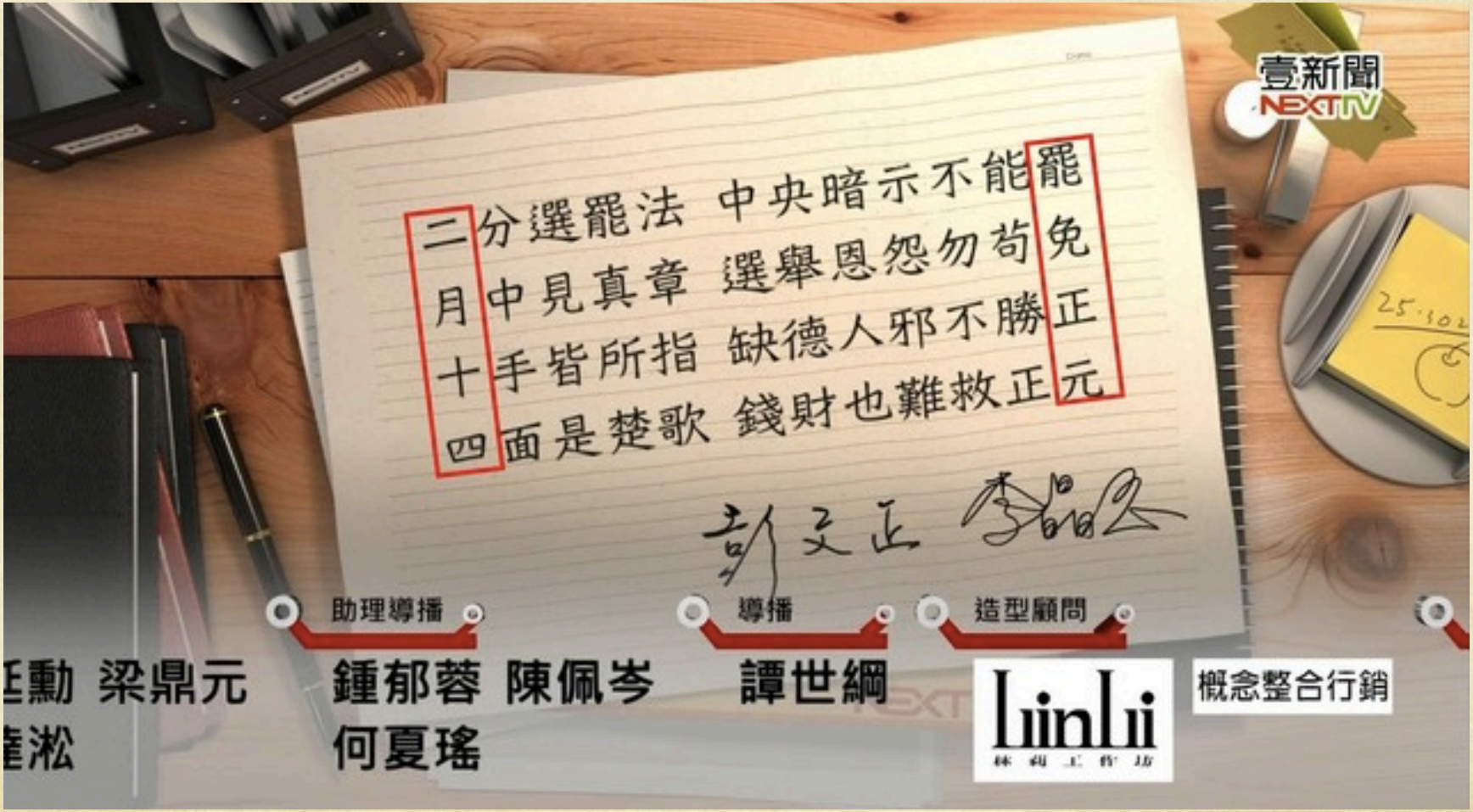
by Mark Chang

- 動機
- 目前演算法
- 後續改良方法

動機

http://antiintelgather.github.io/

罷免不能宣傳


三分選罷法 中央暗示不能罷
月中見真章 選舉恩怨勿苟免
十手皆所指 缺德人邪不勝正
四面是楚歌 錢財也難救正元
壹新聞 NEXTTV
助理導播
導播
造型顧問
鍾郁蓉 陳佩岑
何夏瑤
譚世綱
概念整合行銷

藏頭詩產生器

罷無盡日月明主
免教人間萬古人
過江南來何時見
半是不自知出門
止天上青天明月
兀江上人自有分
再來江山東風景
見不見明月夜深

http://poem.kxgen.net/

搜尋

祭

祭止兀
祭禮之雨
祭禮東衣
祭十二郎文
祭祀公業
祭祀公業條例
祭改
祭文
祭司之林
祭血衚

蒐集語料庫

- 語料庫(corpus)：
  - 語言學上意指大量的文本，通常經過整理，具有既定格式與標記
- 蒐集全唐詩兩萬首作為語料庫

找出句子

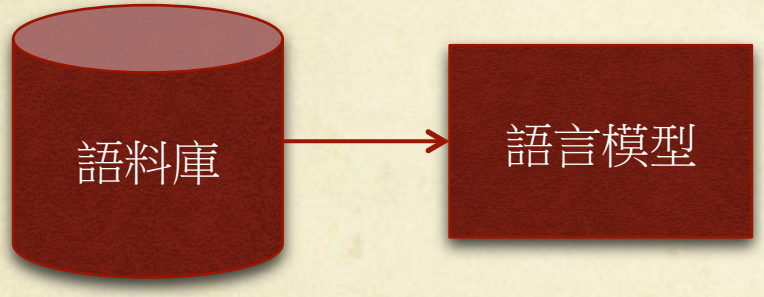
- 直接從語料庫中，找出符合條件的句子
  - 罷 -> 罷唱離歌說遠遊
  - 免 -> 免令仙犬吠劉郎
  - 過 -> 過盡千帆皆不是
  - 半 -> 半夜潛身入洞房

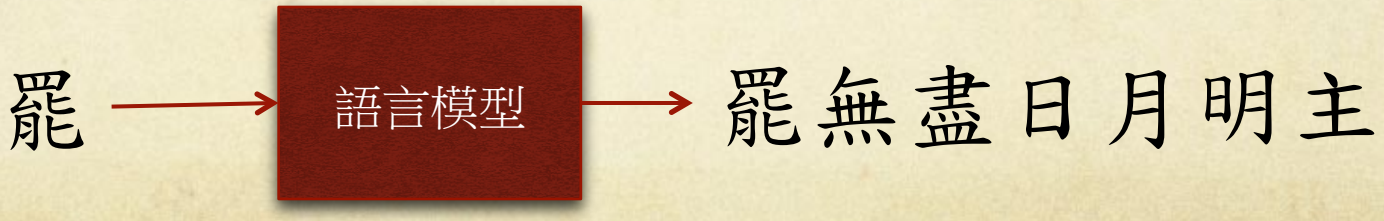
缺點

- 句子和句子之間無關聯性
- 可能找不到某字剛好在某個位置
- 無變化性
- 抄襲！！

目前演算法

- 從語料庫建立語言模型(Bigram)
- 從語言模型算出比較有可能出現的句子(Viterbi)





語言模型

床前明月光，疑是地上霜。
舉頭望明月，低頭思故鄉。

- Unigram
  - 床|前|明|月|光|疑|是 ....
  - 頭：2、床：1 ....
- Bigram
  - 床前|前明|明月|月光 ....
  - 明月：2、地上：1 ....

比較有可能出現的句子

- 若一個句子有五個字，任意挑選五個字，則這五個字符合語言模型的程度，可用機率來表示：
- $P(X_1=w_1, X_2=w_2, X_3=w_3, X_4=w_4, X_5=w_5)$

$= P(X_1=w_1) \times P(X_2=w_2|X_1=w_1)$

$\times P(X_3=w_3| X_2=w_2, X_1=w_1)$

$\times P(X_4=w_4| X_3=w_3, X_2=w_2, X_1=w_1)$

$\times P(X_5=w_5| X_4=w_4, X_3=w_3, X_2=w_2, X_1=w_1)$

比較有可能出現的句子

- Markov Assumption：
  - 每個字出現的機率，只跟前一個字有關
- $P(X_1=w_1, X_2=w_2, X_3=w_3, X_4=w_4, X_5=w_5)$

$= P(X_1=w_1) \times P(X_2=w_2 \mid X_1=w_1) \times P(X_3=w_3 \mid X_2=w_2) \times$

$P(X_4=w_4 \mid X_3=w_3) \times P(X_5=w_5 \mid X_4=w_4)$

- Ex: 白日依山盡

$P(X_1=白, X_2=日, X_3=依, X_4=山, X_5=盡)$

$= P(X_1=白) \times P(X_2=日 \mid X_1=白) \times P(X_3=依 \mid X_2=日) \times$

$P(X_4=山 \mid X_3=依) \times P(X_5=盡 \mid X_4=山)$

比較有可能出現的句子

- $P(X_2 = w_2 \mid X_1 = w_1)$
  - $= P(X_1 = w_1, X_2 = w_2) \ / \ P(X_1 = w_1)$
- $P(X_1 = w_1)$
  - $=$ count(unigram(w_1)) / count(all unigrams)
- $P(X_1 = w_1, X_2 = w_2)$
  - $=$ count(bigram(w_1, w_2)) / count(all bigrams)

產生句子

- 給定句子中的首字w_1,，找出其他四個字：w_2, w_3, w_4, w_5
- 算出使條件機率為最大值的 w_2, w_3, w_4, w_5
  - $P(X_1=w_1, X_2=w_2, X_3=w_3, X_4=w_4, X_5=w_5 \mid X_1=w_1)$

  $=P(X_1=w_1, X_2=w_2, X_3=w_3, X_4=w_4, X_5=w_5) / P(X_1=w_1)$

  $=P(X_2=w_2 \mid X_1=w_1) \times P(X_3=w_3 \mid X_2=w_2) \times P(X_4=w_4 \mid X_3=w_3) \times P(X_5=w_5 \mid X_4=w_4)$
- 給定首字「罷」，產生另外四個字
  - $P(X_1=罷, X_2=w_2, X_3=w_3, X_4=w_4, X_5=w_5 \mid X_1=罷)$

  $= P(X_2=w_2 \mid X_1=罷) \times P(X_3=w_3 \mid X_2=w_2) \times P(X_4=w_4 \mid X_3=w_3) \times P(X_5=w_5 \mid X_4=w_4)$

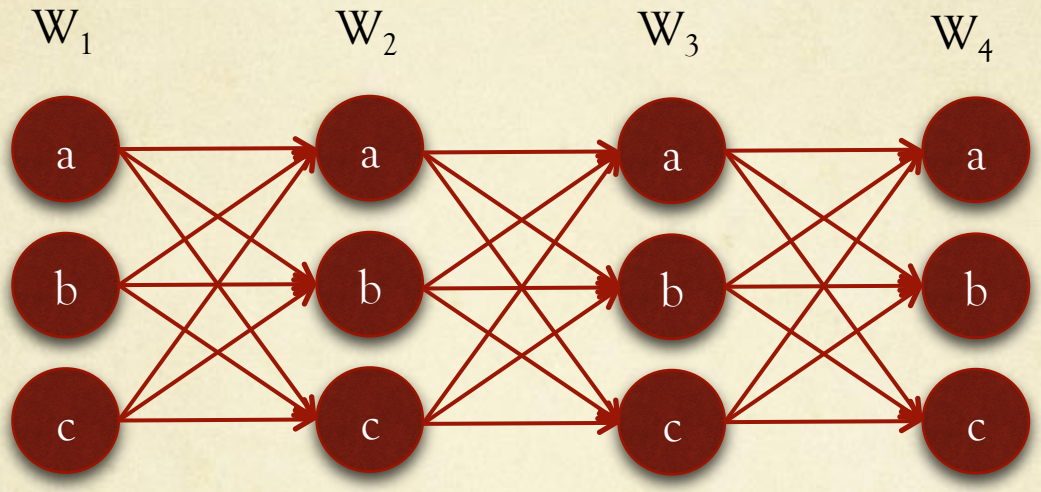
Bigram Smoothing

- 若在Bigram中未出現w_1, w_2組合
  - 則$P(X_1=w_1, X_2=w_2) = 0$，會使得整個句子機率為0
- 為了避免此現象發生
  - 令$P(X_1=w_1, X_2=w_2) = 0.5 / \text{count(all bigrams)}$

時間複雜度

- 給定w_1,，求機率最大值的w_2, w_3, w_4, w_5
- 若詞庫中有3000個字可挑選，則共有3000^4種組合
- 若詞庫大小為V，句子長度為L，則時間複雜度為$O(V^L)$

時間複雜度

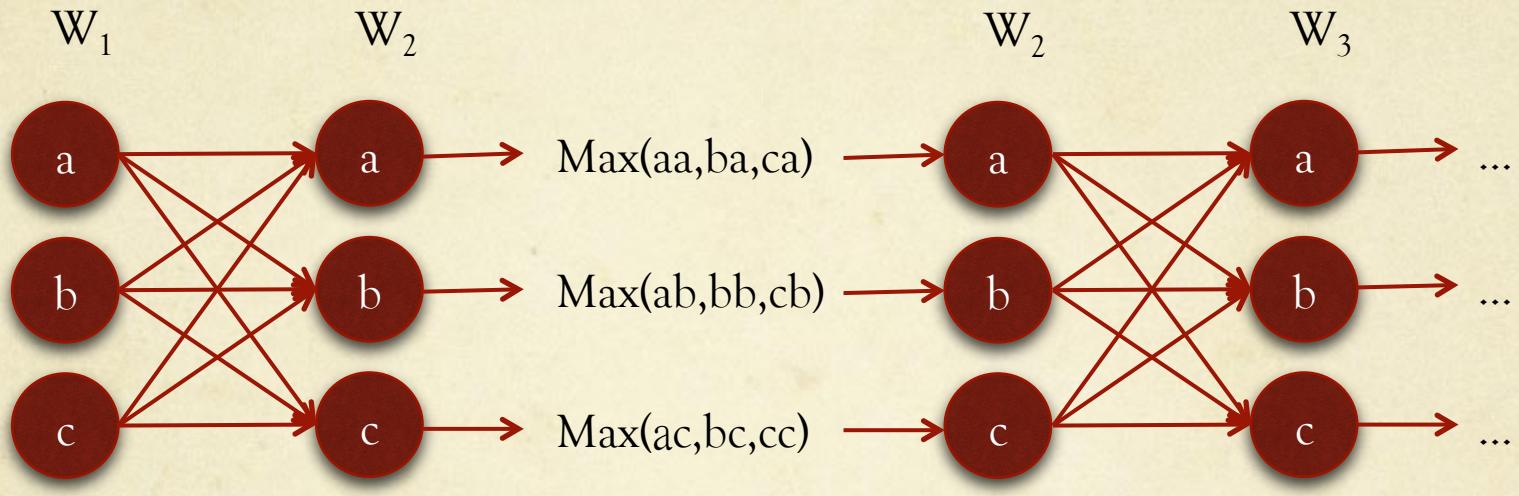


- Max(aaaa, aaab, aaac, aaad, aaba, aabb, aabc ...)
- 共 V^L 總組合，求其中的最大值。

Viterbi演算法

- Dynamic Programming
- 不需要先把所有組合算出來，再求最大值
- 先算局部的最大值，傳遞下去
- 若詞庫大小為V，句子長度為L，則時間複雜度為$O(LV^2)$

Viterbi演算法



- 先算 Max(aa, ba, ca) ，只保留最大值傳遞下去
- 每層(W_{i-1},W_i)需計算 V×V次
- 總共需計算 V×V×(L-1)次

計算結果

罷無盡日月明主
免教人間萬古人
過江南來何時見
半是不自知出門
止天上青天明月
兀江上人自有分
再來江山東風景
見不見明月夜深

缺點

- 語句不夠通順
- 句子和句子之間無關聯性

源碼

- Python版：
  - https://github.com/ckmarkoh/AcrosticPoem
- Javascript版：
  - https://github.com/ckmarkoh/AcrosticPoemJS

後續改良方法

- Machine Translation
  - Phrase-based SMT Model
- Neural Network
  - Neural Probabilistic Language Model
  - Multiplicative Recurrent Neural Network

Phrase-based SMT Model

- Ming Zhoua, Long Jianga, and Jing Heb. Generating Chinese Couplets and Quatrain Using a Statistical Approach
- DEMO：http://couplet.msra.cn/app/couplet.aspx

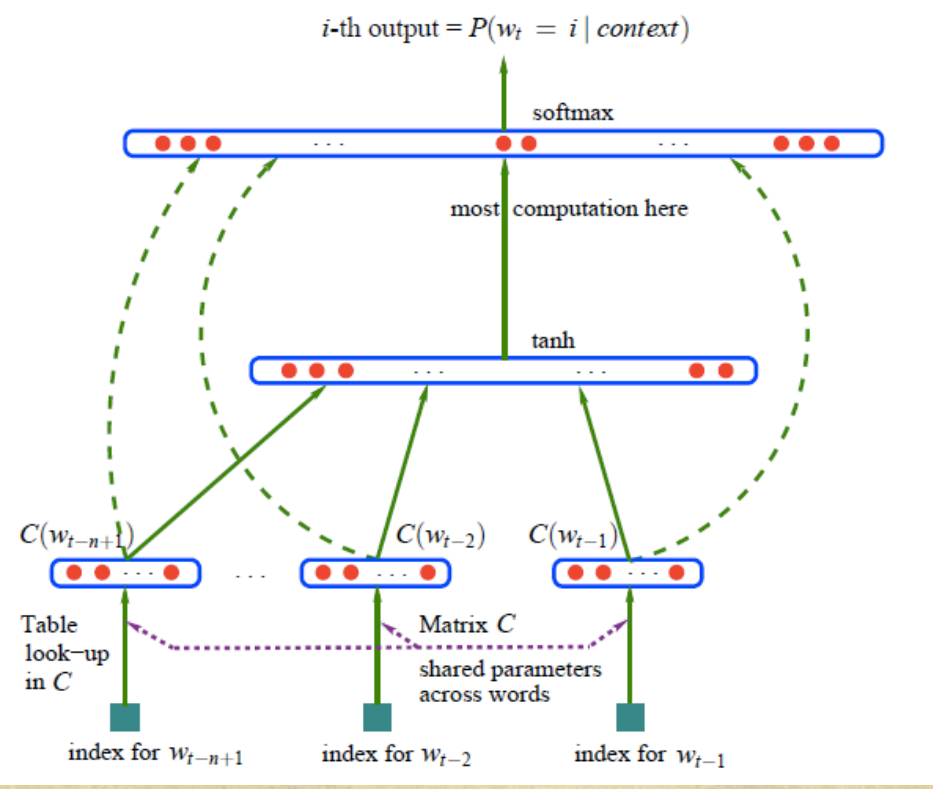
FS:	海(hai) sea	阔(kuo) wide	凭(pin) allow	鱼(yu) fish	跃(yue) Jump
	|	|	|	|	|
SS:	天(tian) sky	高(gao) high	任(ren) permit	鸟(niao) bird	飞(fei) fly

$$S^* = \arg\max_S p(S \mid F)$$

$$= \arg\max_S \sum_{i=1}^{M} \lambda_i \log h_i(S, F)$$

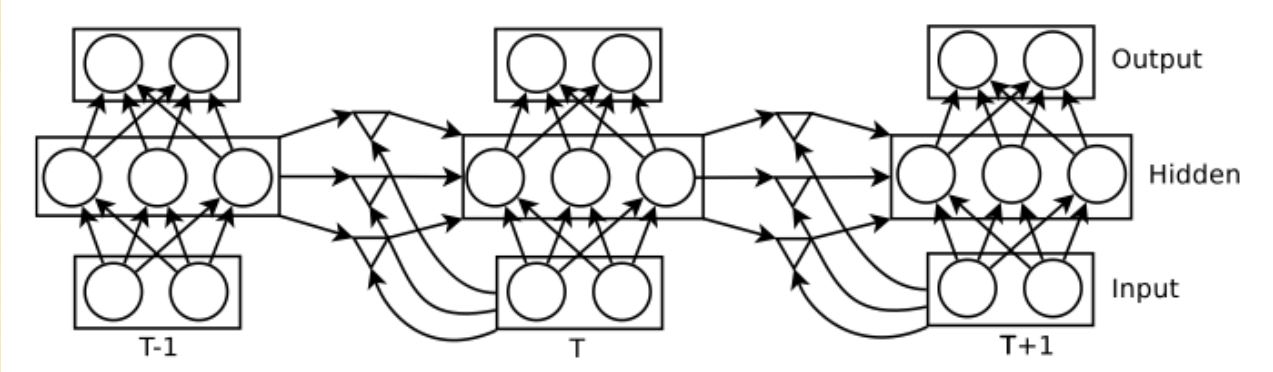
Neural Probabilistic Language Model

- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin. A Neural Probabilistic Language Model



Multiplicative Recurrent Neural Network

- Ilya Sutskever, James Martens, Geoffrey Hinton. Generating Text with Recurrent Neural Networks
- DEMO： http://www.cs.toronto.edu/~ilya/rnn.html



$$f_t = \text{diag}(W_{fx}x_t) \cdot W_{fh}h_{t-1} \quad (7)$$

$$h_t = \tanh(W_{hf}f_t + W_{hx}x_t) \quad (8)$$

$$o_t = W_{oh}h_t + b_o \quad (9)$$

The End