

文字檔案存取、JSON格式及 句法學

台師大通識教育課程

文本分析與程式設計

授課：卓騰語言科技 _ PeterWolf

句法是二元樹的時代了！

- 透過「二元」的句法樹(Syntax Tree), 單一個句法樹, 就能處理人類全部的自然語言 (還記得「所有的人類都具有一樣的語言處理機制」嗎?)

<https://medium.com/人工智慧/語法理論-23bc87126e6>

M | 人工智慧

語法理論



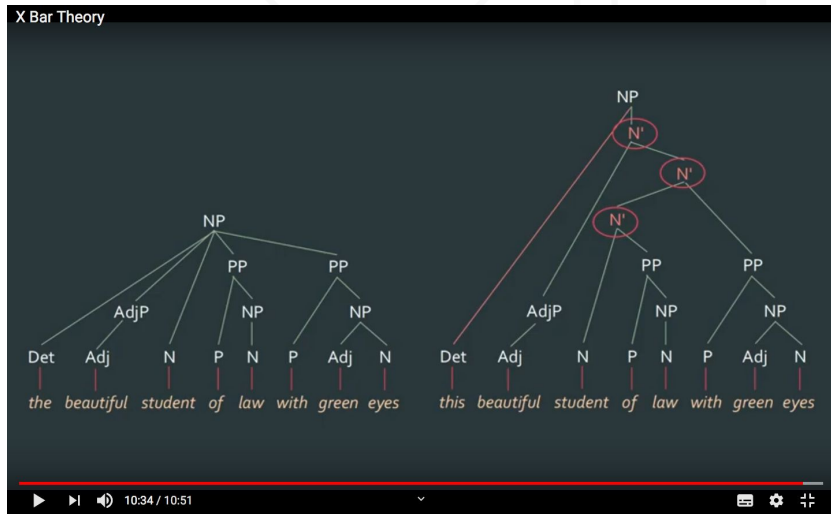
陳鍾誠 Follow

Mar 5, 2019 · 2 min read

BNF/EBNF

- 以下圖片來自陳鍾誠的語言處理技術一書 —

<https://www.slideshare.net/ccckmit/ss-15898210>



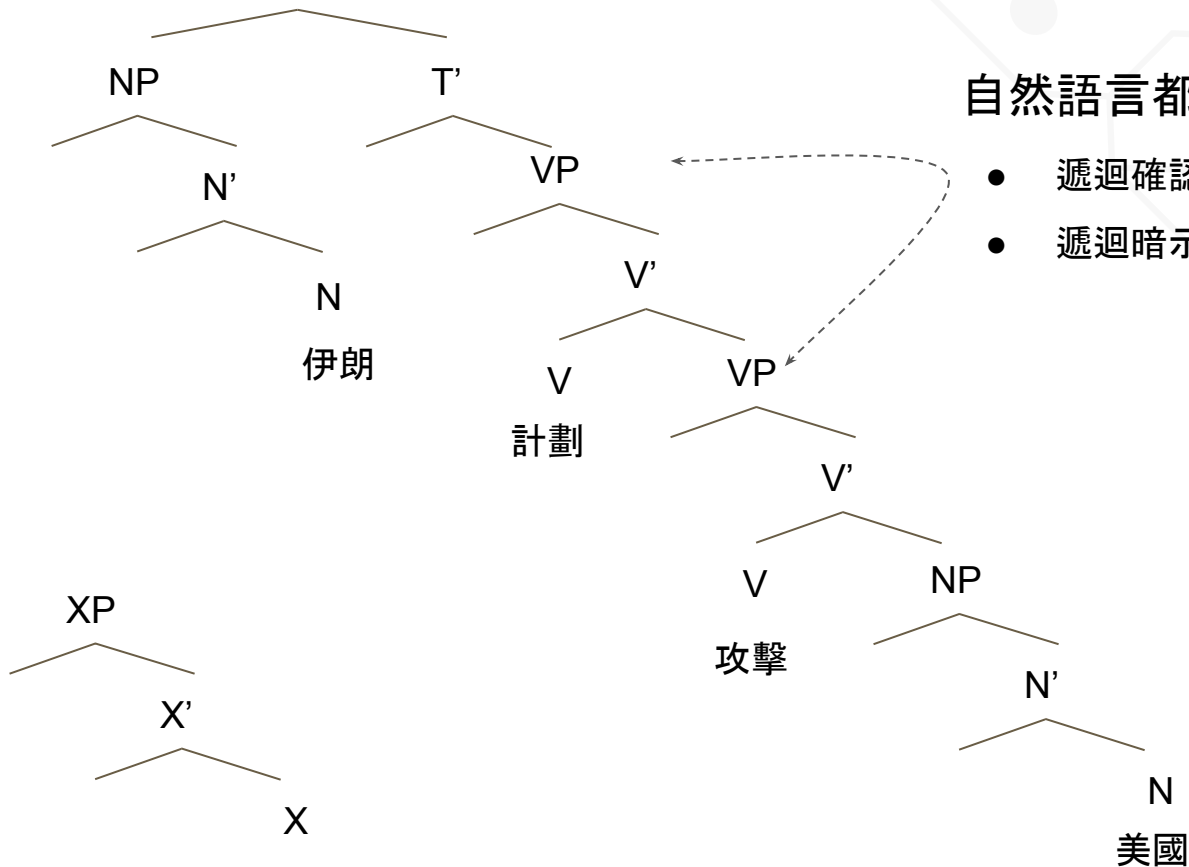
<https://youtu.be/FNvua42KZRw>

NLP 技術的發展限制



<https://pdfs.semanticscholar.org/7b2f/6eb72fe47a772a6c51acaeff34071cb596ba.pdf>

伊朗計劃攻擊美國



自然語言都有「遞迴」的結構。

- 遞迴確認了一件事：世界上不存在最長的句子。
- 遞迴暗示了一件事：句子具有最小結構

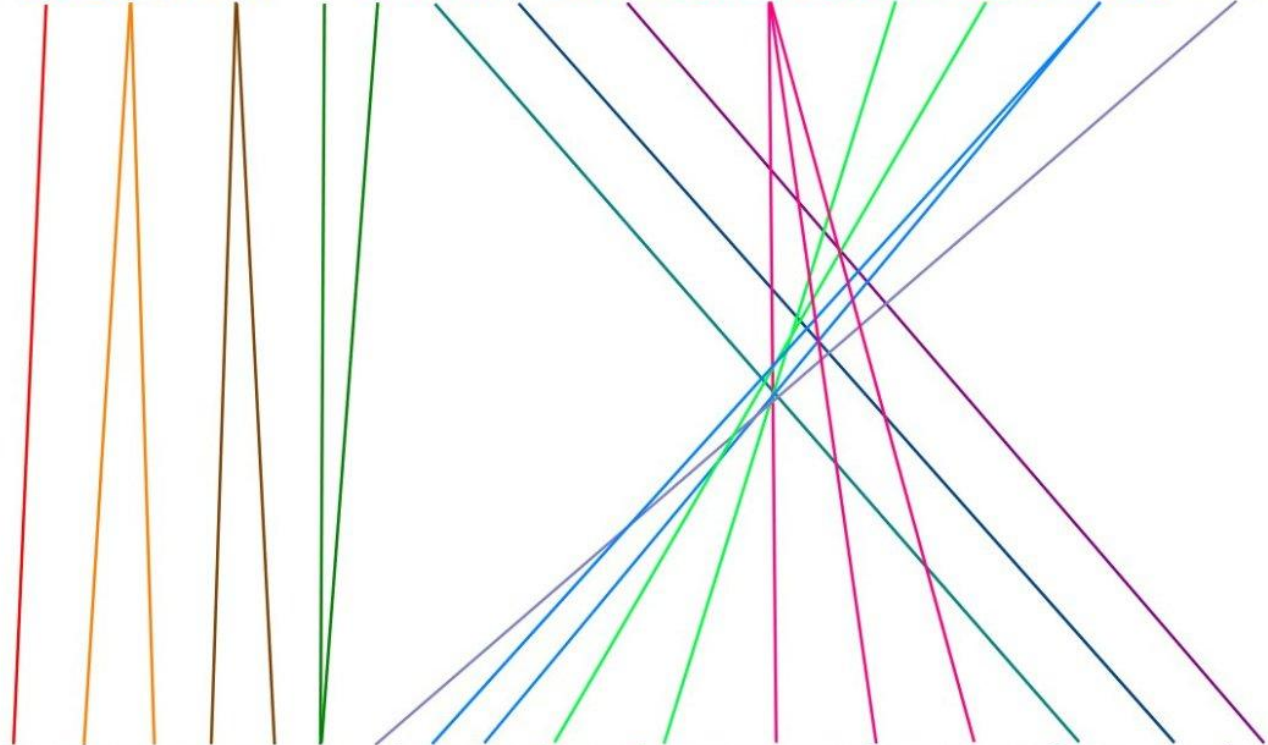
NLP 技術的發展妄想

<https://awed-frog.tumblr.com/post/617561392469393408/whoa>

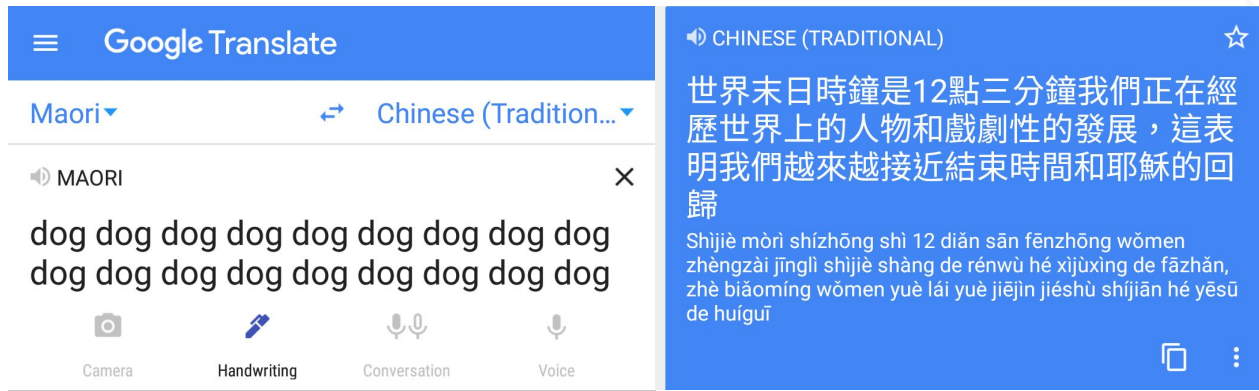
Chinese: 我想要試穿一套從我們飯店對面的商店裡看到的西裝。

u/stupidegg_

English: I want to try on a suit I saw in a shop across the street from our hotel.



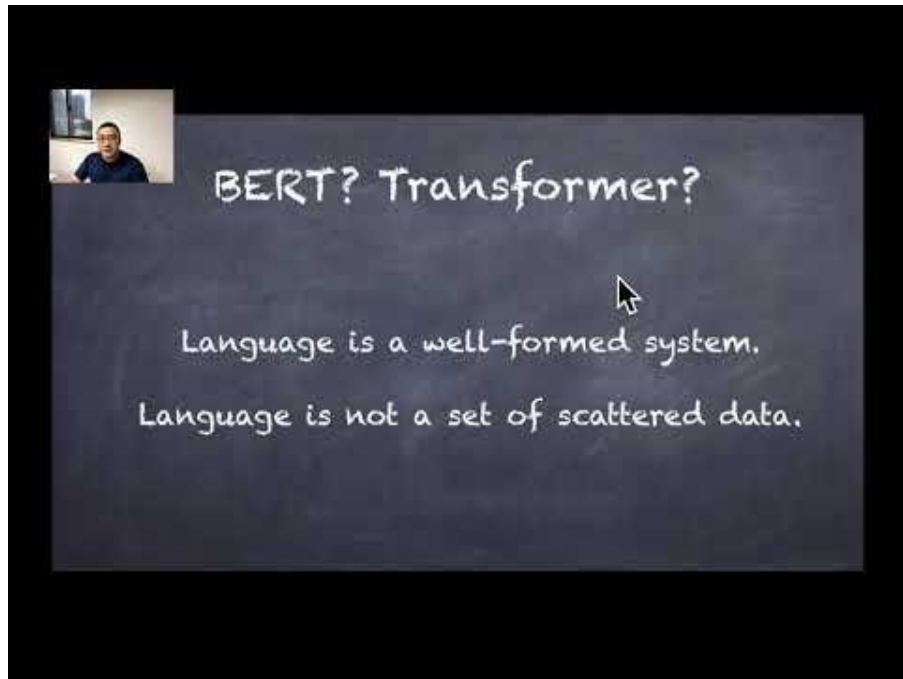
當 NLP 只剩下統計和機率的觀點時...



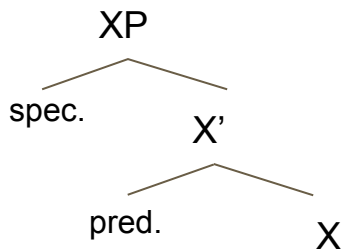
<https://www.thenewslens.com/article/100431>

One Tree, One Mind, One Species

<https://youtu.be/6Jq-4Pu3er8>

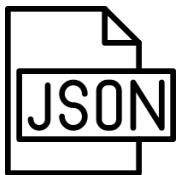


Insight of binary tree: Key-to-Value



JSON 格式全名叫做 JavaScript Object Notation, 用**大括號 {}** 包住所有的元素, 每個元素用 **半型逗號 ,** 分開

```
{  
    key: value  
}
```



XP {

}
}

X' : [____, x],
spec : ____

X' {

}
}

X : ____
pred : ____

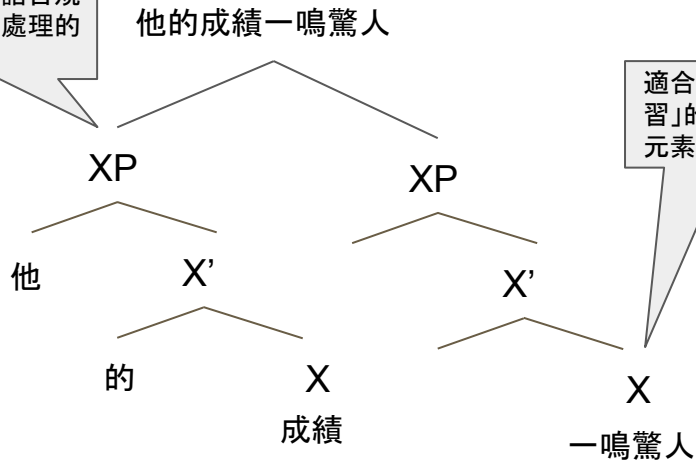
Quiz: 回想至今三堂課所學，試回答以下問題：

1. 請思考以下的三棵句法樹，哪一棵有道理？

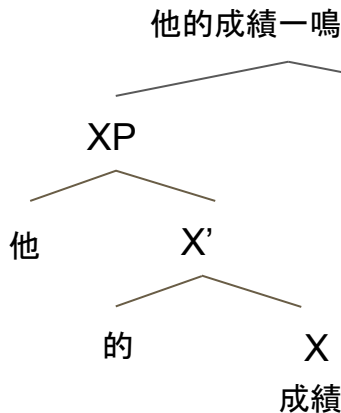
成語

成語是東亞特有的語言形式，存在於在漢語、日本語、朝鮮語、越南語等語言中，它是一個固定短語，表達了一個固定的語意，常帶有歷史故事及哲學意義。成語大多數由四個漢字組成，在文言文中使用廣泛，也有其他字數不等的成語，在白話文或者日常會話中也經常出現。 [維基百科](#)

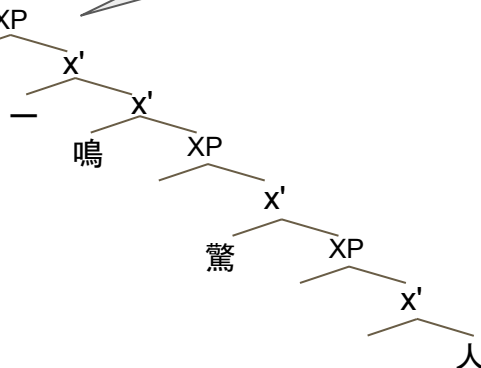
適合利用「語言規則」的方式處理的元素。



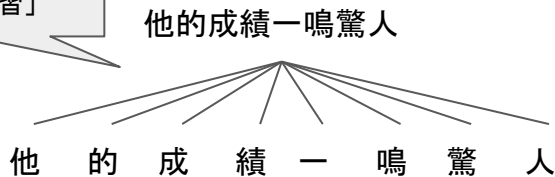
適合利用「機器學習」的方式處理的元素。



過度利用「語言規則」



過度利用「機器學習」



Assignment:

1. 從課程 github repo 中把課程中提供的 week03 的目錄載下來。
2. 把 week03.py 改名為 **week03_你的學號.py**
3. 在 **week03_你的學號.py** 中，設計你的程式完成以下功能：
 - a. 讀取 example.txt 中的文字，並印出在畫面上
 - b. 仿照 week03_example.json 的格式，另外新增一個名為 **week03_你的學號.json** 的檔案，並將 example.txt 中的內容，利用 json 模組的 dump() 寫入 **week03_你的學號.json** 中
 - c. 讀取 **week03_你的學號.json** 的內容，並印出在畫面上

git 初學者的四個動作...

0. `git clone xxx.git` #複製遠端的 .git 源碼倉 (repository) 到本機

1. `git add xxx` #加入 xxx 準備操作它

本

2. `git commit -m "ooo"` #說明你對 xxx 做了 ooo 的操作

3. `git pull` #把最新的源碼庫版本抓下來 (取得更新權)

遠

4. `git push` #把你的 xxx 傳上去

5. `git rm xxx` #移除 xxx (之後仍要跑 commit, pull, push)