http://bit.ly/rocling2020cwn

ROCLING Tutorial on Sense-aware computing in Chinese

中文計算語意入門

謝舒凱 & 曾昱翔　台大語言學研究所　2020/09/26

講者介紹

謝舒凱

台大語言學研究所副教授/所長

台大人工智慧與機器人研究中心

台大人文創新與全球化中心

台大神經生物與認知科學中心

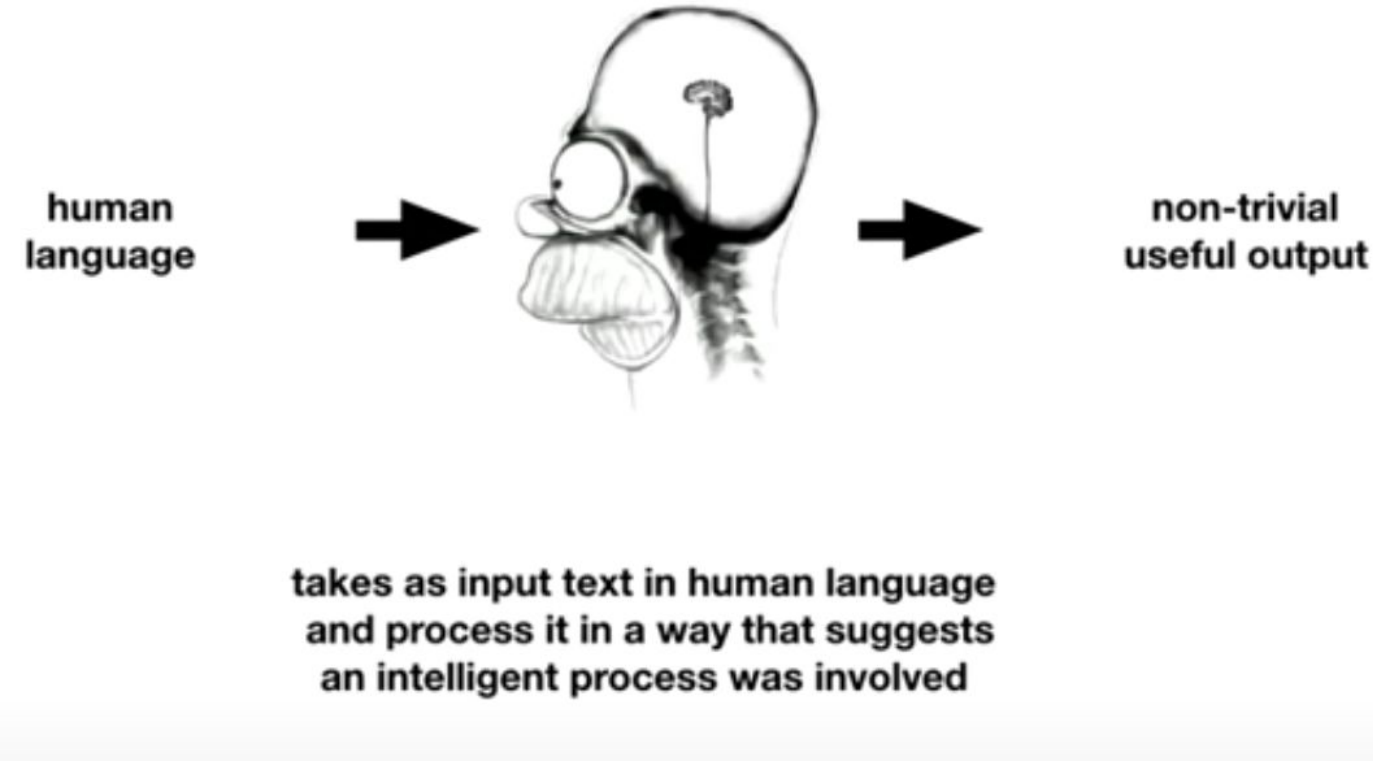
曾昱翔

台大語言學研究所AI 創新研究中心專案博士後研究

議程介紹

11:00-11:40 簡介 | Introduction to semantic computing

11:45-12:30 上手實作 | Hands-on session



NLP in a nutshell (Yoav Goldberg, 2019)

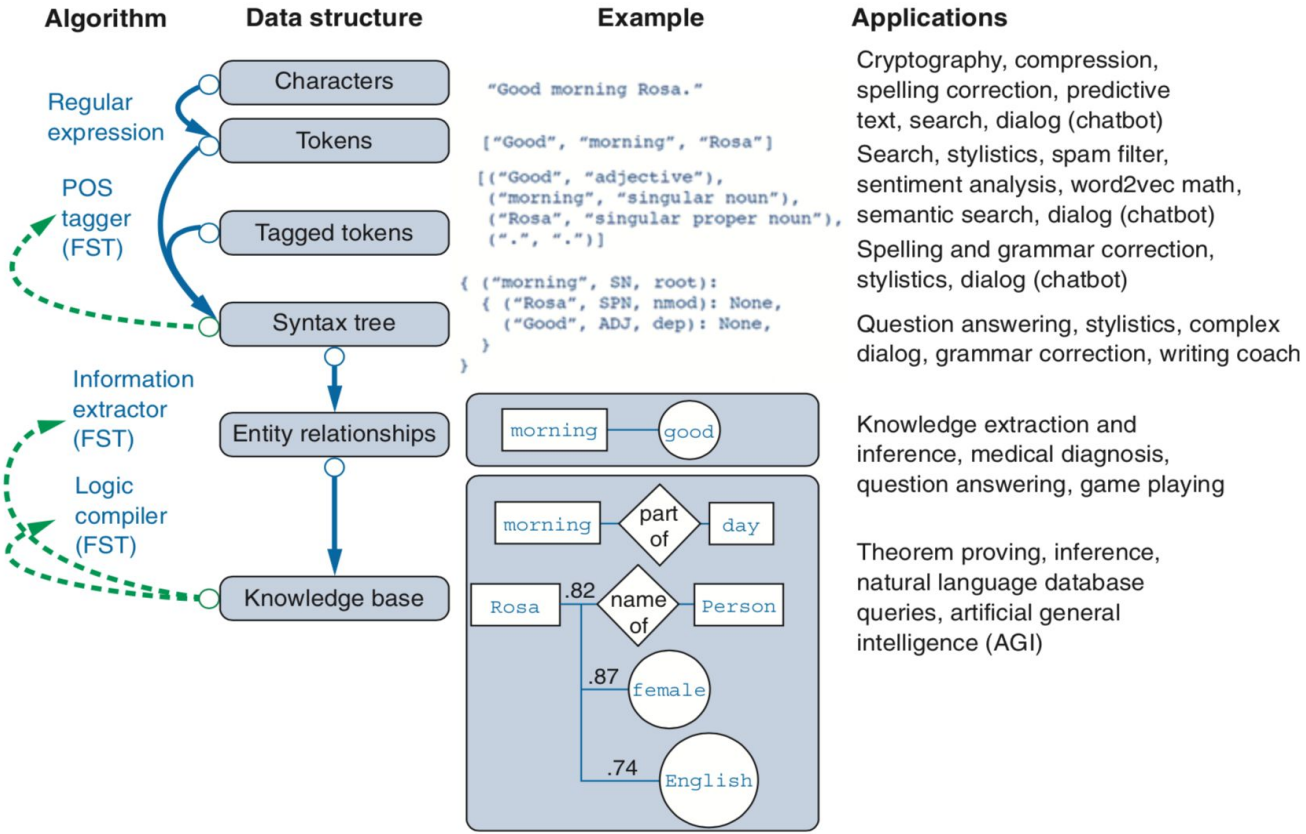
NLP in Action

其中的【理解】機制是最不好理解、不好定義的

- **NLP: Understanding**, analyzing and generating language/text

我們大概知道理解是跟意義的【表徵與處理】有關

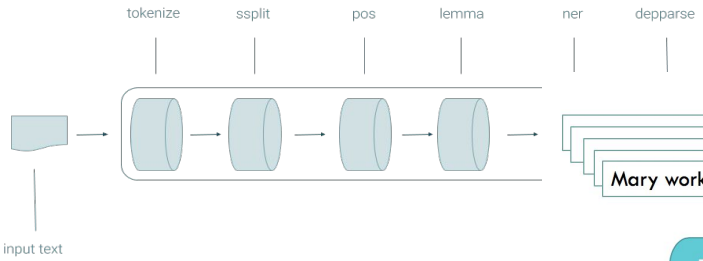
- Meaning representation and manipulation *is* the core

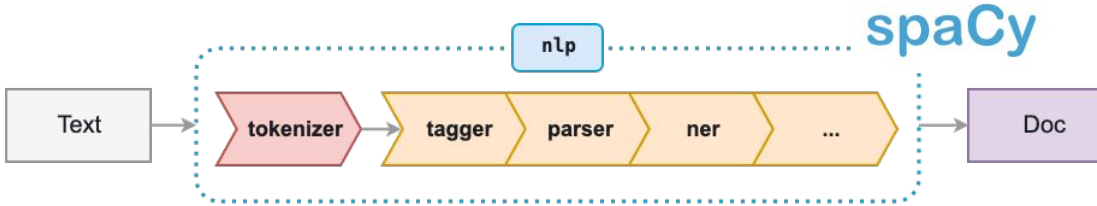


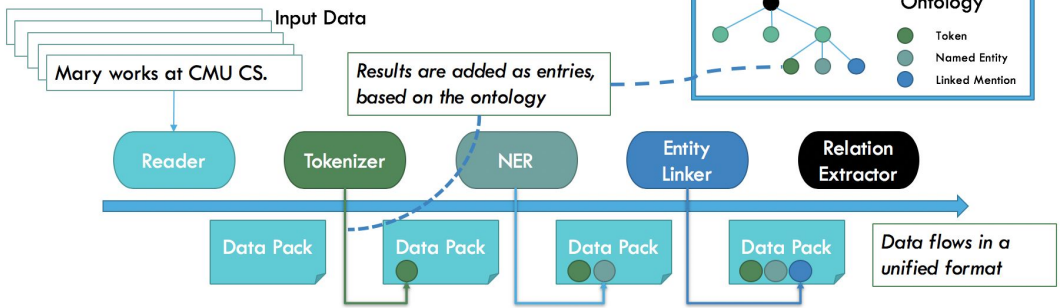
NLP Pipelined architecture

Natural Language Processing depth (Lane et al. 2019)

百家爭鳴的管線架構 | Stanford coreNLP, Spacy, Forte, etc







Forte

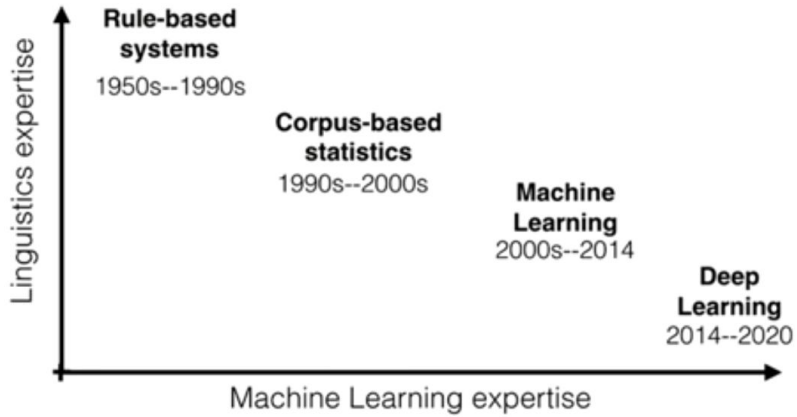
缺了重要的一塊 | a missing piece in NLP (pipeline)

詞義自動消歧 (Word Sense Disambiguation) 任務。接在管線架構時可以稱作**詞義自動標記系統** (sense tagger)。

Deep Learning > NLP

From labor-intensive and data-light approach to (the opposite) ?

Distributed (vectorized semantic) representation is the key factor



語意學、語言理解與 AI | semantics, language understanding and AI

語言理解的【操作型定義】

- 把剖析結果用人可以詮釋的樣子呈現出來 | Modularized, Pipelined Natural Language **Processing** yields (partial) symbolic representation of the results.
- 用任務來確認 | Extrinsic evaluation / Task-oriented validation (in terms of Turing test?)

Baby steps with reasoning

但是我們期待的理解能力還包括了推理、類比、隱喻運算

用邏輯蘊涵為例 | entailment as example (邏輯表徵不夠之處)

我們怎麼進行推理的(邏輯+ 語言知識)

(a) Mary is tall and thin.(*premise/antecedent*)

(b) Mary is thin.(*conclusion/consequence*)

(a) *entails (b), denote it* $(a) \Rightarrow (b)$

(a) Sue only drank half a glass of wine ⇒ Sue drank less than one glass of wine.(*measures and quantity*)

(b) A dog entered the room ⇒ An animal entered the room. (*word meaning relations*)

(c) John picked a blue card from the pack ⇒ John picked a card from the pack.(*adjective modification*)

但還有世界知識、語用(大魔王) | World knowledge and Pragmatics

(defeasible reasoning)

(a) Tina is a bird.

(b) Tina can fly.

(c) Tina is a bird, but she cannot fly, because ... (she is too young to fly, a penguin, etc)

那, (人 - 機) 需要什麼樣的語意學 | which semantics

先問大家「電腦」是什麼意思

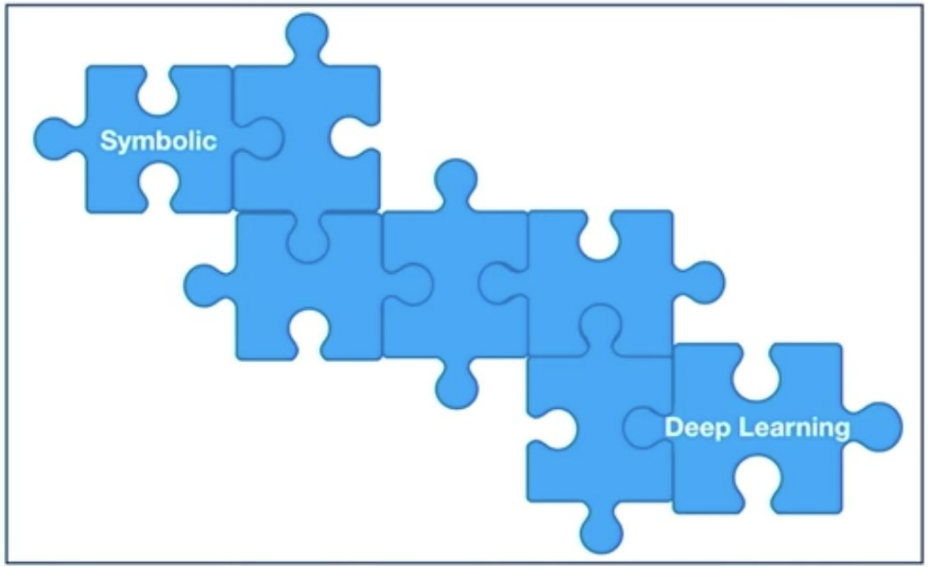
Frege 的洞見 : Sinne (sense) and Bedeutung (denotation)

#1	事務所 給 我 登記 的 是 「 王 + 肉字邊 」 (	電腦	無法 顯示 這 個 字) 導致 我 護照 通關 差點 有
#2	沒 刻好 好 可愛 喔 ！ 請問 是 哪 間 呢 ？	電腦	字 + 蓋 的 不 好 … （ 小聲 說 : 這
#3	瀏覽器 完成 購物 下單 有 領 過 回饋 不 放心 可 用	電腦	版裝 line 購物 賺 點 小 幫手 開 網頁 時 會 提醒
#4	老王 嗎 ？ 他 理由 是 醫師 背對 他 看 不 到	電腦	螢幕 ？ 明明 牆壁 有 電視 可 看 ！ 跟 老公 溝通
#5	我 跟 我 先生 有 進去 客房 （ 兼 書房 ） 用	電腦	， 沾到 更多 病毒 ， 結果 我們 週末 發病 ， 半夜 吐
#6	耐力 嚴 重 不足 . 後天 大 近視 . 3 c 手機	電腦	兒童 到時候 天使 長輩 就 稱不上 天使 了 ， 除非 你 夠
#7	， 8 坪 和 10 坪 的 差別 。 其他 就是 靠	電腦	的 簡報 在 介紹 。 (後來 得知 房型 分 3 種
#8	政府 承辦人 ， 他們 說 不會 被 關說 ， 現在 都 是	電腦	抽籤 。 但 我 還是 希望 把 這 問題 提出來 ， 大家

人是符碼動物 | symbolic species

理解(meaning representation and manipulation)

最後應該還是要回到解釋與行動

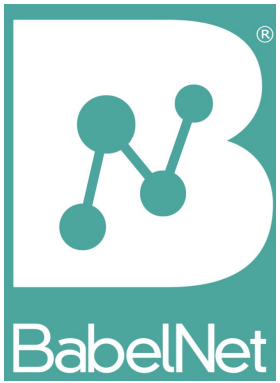


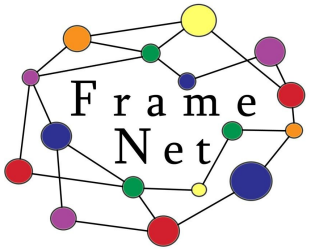
計算詞彙語意學 | Computational Lexical Semantics

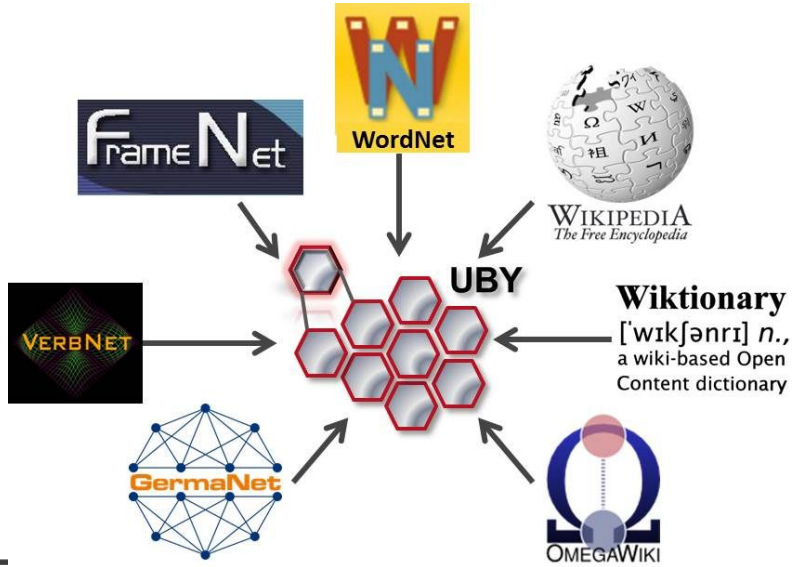
應該是一個不錯的選擇方案

- 看看語言系統怎麼在各種語境中【使用】| language in use
- 詞義、關係、知識本體 | word sense, relations, ontology

各種詞彙語意知識資源 | lexical semantic resources getting real







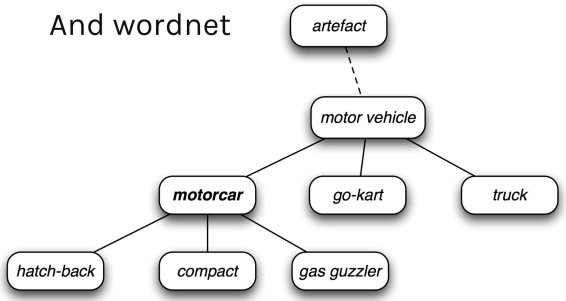
yago
select knowledge

WordNet

- 同義詞集 | Synset (synonymous set) as concept
- 詞彙關係構成概念網路 | Synsets are interconnected via different lexical (semantic) relations
- 全球連結 | Global WordNet Grid
- 開放、可擴充（情緒、多模態、事實訊息）

python™
Natural Language Analyses with NLTK

And wordnet



Chinese Wordnet https://lopentu.github.io/CwnWeb/

- 考量詞頻的詞彙
- 詞彙語意理論
  - 多義行為與詞義粒度、釋義控制詞彙、詞彙關係分類判準
- 跨語言連結
- (2.0) 心理實驗資料驗證與互動、詞義與關係數量擴增、人工標記詞義語料庫、自動詞義標記、檢索與程式介面

中文詞彙網路
CHINESE WORDNET

最新消息

📢	歡迎使用 CWN2.0 線上查詢！	*2020/08/08*
📢	歡迎參與CWN詞意標記工作小組！	*2020/07/01*
📢	CWN2.0 Python程式介面已釋出	*2020/05/21*

CwnGraph Github　直接下載 CWN 資料

12,753	34,422	12,620	47,450
詞條	詞意	同義詞集	詞意關係

關於 中文詞彙網路

CWN 簡介　相關人員　研究發表

CWN 簡介

線上查詢

吃

1. **使物體經過口中吞入體內。** VC
一隻猴子會分辨什麼果子能<吃>，什麼果子不能<吃>，這屬
同義詞 用 食 啖 啃 進 喀

2. **服用藥物。** VC
練氣功，也像<吃>藥一樣，各種功法，對不同的經絡有不同

3. **用牙齒磨咬物品。** VC
你眼所見、腳所踏都是珍貴藝術品，所以不可觸摸物品，在
同義詞 嚼 咬

4. **將物體含咬在口中，為不自主的習慣動作。** VC
他怎麼會<吃>手指頭、<吃>筆啊？他是不是有點腦袋有問題
同義詞 咬 啃
下位詞 吸 吸食

5. **使用會令人上癮的物品，通常用口或鼻攝入。** VC
嫌犯於警訊時，矢口否認餵兒子<吃>強力膠，強調他是在「
同義詞 吃下 吃下去 吸 吸食
下位詞 抽

6. **偏好後述口味或烹調方式。** VC
可能是因為最近<吃>得很淡只要我一<吃>油的東西，我就想

7. **在後述地點用餐。** VC
這個月光<吃>館子的錢就已經花了我三千多塊了！

8. **比喻身體遭受後述事物的攻擊。** VC
他<吃>了三顆子彈，兩顆在前胸，一顆在下腹。
同義詞 吃下去 挨 吃下

9. **比喻經歷後述負面事件。** VC
說得誇張一點，過去一年來，<吃>官司的銀行董事長或總經
同義詞 吃下去 吃下

10. **比喻佔便宜。** VC
劉若英被<吃>豆腐、遇色狼的經驗太多了，公司只好請私人

11. **比喻依靠後述對象過生活。** VC
連頒獎、晉級也沒份，記功也只有一個小功，你們知道的，

12. **比喻物品或能量因使用而漸漸減少。** VC
在農產品的低利潤之下，「堅持」可能變成一種<吃>本錢度
同義詞 消 出 消耗

13. **比喻以不正當的方式取得財物。** VC
從健保卡「總歸戶」的工作中，還可以連帶揪出和病人勾結
同義詞 吃下去 黑 歪 污 吃下 汙

14. **比喻絞到後述細長物體。** VC
使用吹風機時，吹風口不要離頭髮太近，以免<吃>到頭髮。
同義詞 咬

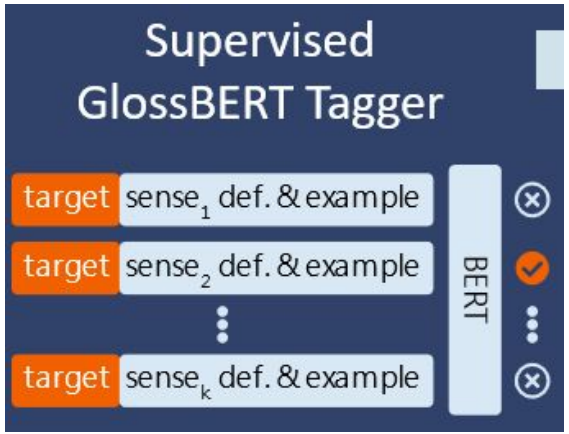
15. **比喻爭取到後述好處。** VC
物價蠢動，糖品業者猛囤貨，搶<吃>價差的甜頭。

WN-based WSD

- WSD ⟪⟫ Sense tagger
- **SemCor:** Sense-tagged corpus

詞意消歧模型

- 透過人工標記以及CWN中既有例句，我們透過GlossBERT建立詞義消歧模型。
- 在該架構中，模型學習的是一組句子/詞意配對是否正確：
  - 依雜誌所<言>/言,以文字媒介引述或陳述訊息,西諺常<言>
  - 依雜誌所<言>/言,計算中國詩中每一句固定字數的單位,分為四<言>古詩



GlossBERT (Huang et al., 2019)

BERT, GlossBERT, DistillBERT

The resulting WSD model, which is based on GlossBert (Huang et al, 2019), has accuracy of 83%.

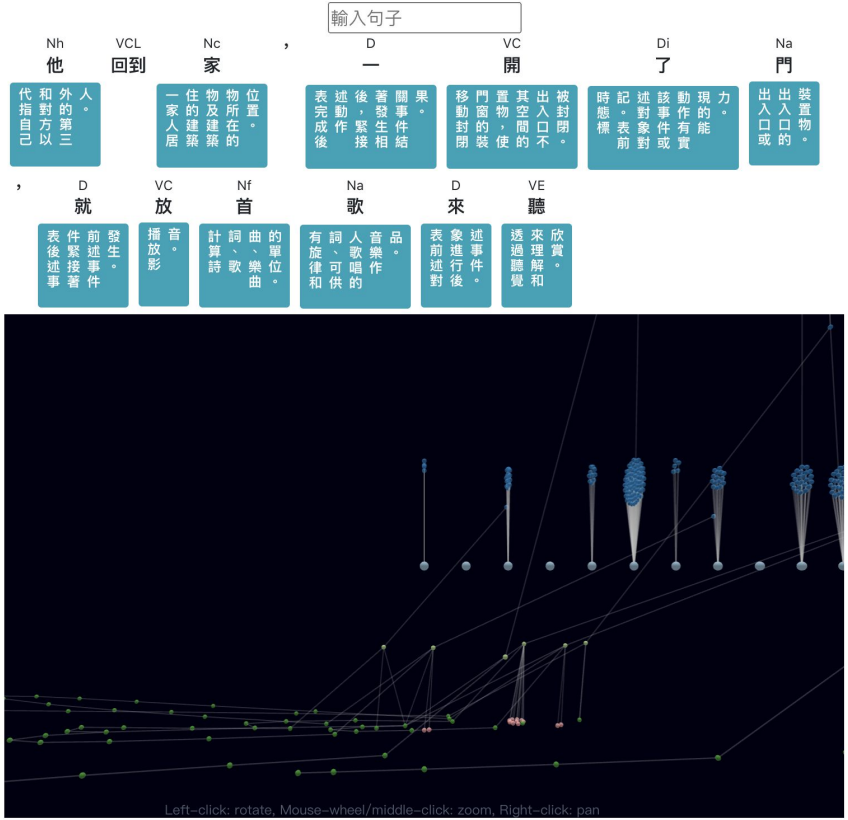
CwnSenseTagger

CwnSenseTagger 0.1.6

```
pip install CwnSenseTagger
```

```
import CwnSenseTagger
txt = "指揮中心籲週末假日非必要別出門"
hard_tag = tagger.tag(txt)
CwnSenseTagger.senseTag(hard_tag)
```

```
[[['指揮', 'Na', '', ''],
  ['中心', 'Nc', '07024204', '具有特定設備，提供特定資源的機構或單位。'],
  ['籲', 'VE', '09249101', '公開表明後述訊息希望得到支持。'],
  ['週末', 'Nd', '03057801', '星期六及星期日。'],
  ['假日', 'Nd', '', ''],
  ['非', 'D', '04038004', '表一定，排除其他可能性。'],
  ['必要', 'VH', '', ''],
  ['別', 'D', '05003501', '表不要。'],
  ['出門', 'VA', '06736201', '從居住地的裡面移動到外面。']]]
```



網路介面 http://140.112.147.132/CwnWsd/

中文自然語言處理環境
CKIP coreNLP

斷詞	|	Word segmentation
詞性標記	|	POS tagging
專有名詞辨識	|	NER
句法剖析	|	Dependency parser

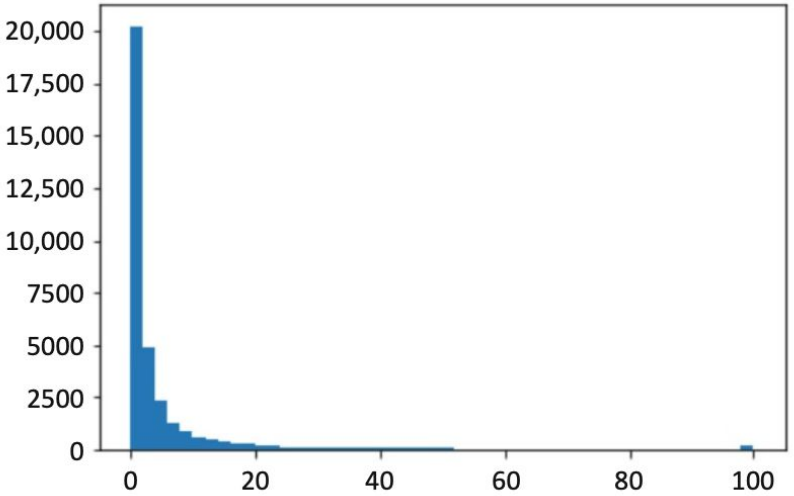
+ 詞義自動標記 | Sense Tagger
+ 人工詞義標記語料庫 (27,012 句) | Chinese SemCor

WSD 能做什麼 | Some thoughts on WSD

- 檢證詞彙語意理論
- Enhanced text mining / data science

詞意頻 (sense frequency)

How often do senses occur in **SemCor**? - 大多數的詞意(20K out of 33K)只出現 1-2 次，只有大約500 個詞意出現超過 100 次以上（語料庫可能太小了）



我們做了較大規模的詞意人工標記

- 選擇了在CWN中113個詞意數超過10個的詞條進行人工標記。
- 詞意標記是以句子為單位。這些句子是從AS平衡語料庫中隨機抽取，每個句子都包含至少一個目標詞。
- 六位標記者經過訓練，以及定期討論，共完成了27,012組詞彙標記。
- 部分人工詞意標記資料可由此連結下載。

Sense Annotation

詞意頻率與詞意向量

- 藉由詞意標記模型，我們對平衡語料庫中約500萬字的語料進行詞意自動標記。藉由這些標記，我們可計算出每個詞意在語料庫中的出現頻率。
- 標記完成之後，每個被標定詞意的詞，都會在該詞之後註明其詞意序號，例如：
  - 「一切(Neqa)　公私(Na)　關係(Na)」
  - 「一切-03049601　公私　關係-05186502」
- 把每個包含詞意序號的詞當成token，我們就可藉由word2vec計算其word (sense) embeddings

詞意頻 | word sense frequency in ASBC

- 從平衡語料庫中取出300 句包含「吃」或「開」的句子（一句可能含多個目標詞），用CwnSenseTagger 判斷該詞在脈絡中的詞意，並計算其頻率。以下列出頻率最高的5 個詞意。

「吃」的詞意	頻率
使物體經過口中吞入體內。	296
服用藥物。	17
比喻經歷後述負面事件。	8
比喻佔便宜。	2
偏好後述口味或烹調方式。	1

「開」的詞意	頻率
形容有花植物的花朵長出並舒展。	32
駕駛後述交通工具。	24
進行會議。	21
創立後述機構。	19
形容物體的部件不緊密結合。	17

Future/Challenges

- 不同詞義粒度的自動標記(coarse and fine-grained sense tagging)
- 規則多義標記Regular polysemy tagging 與 WSD-NER 整合應用
- 詞義層次的syntagmatic + paradigmatic 計算

Wrap up

- 語意計算是 AI-NLP 核心
- 語意表徵 (meaning representation) 是語意計算的基礎。
  - **Sense computing (✓)** vs. denotational computing
  - 符碼表徵給人解讀(symbolic representation for human learning)，向量表徵給機器解讀(distributed representation for machine learning)
- 我們完成了(全球唯一的)中文詞義自動標記系統 + 人工標註語意之語料庫

Python 程式實作

實作介紹

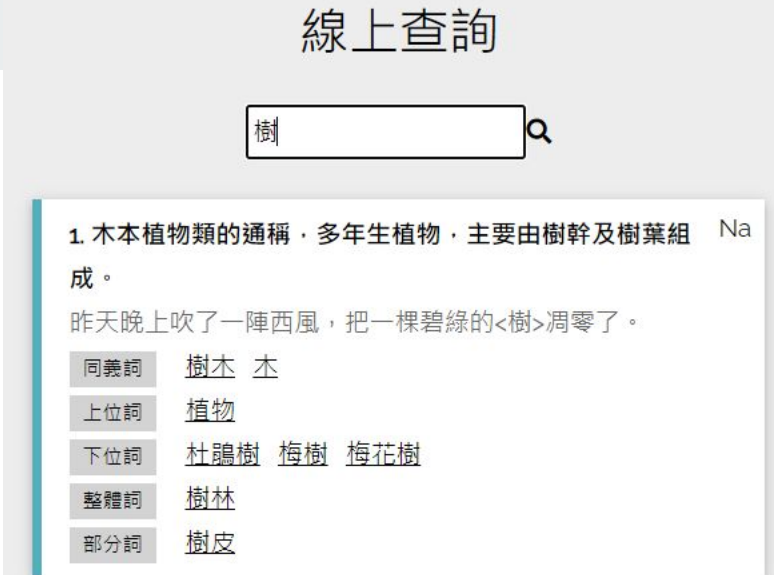
- 我們將會介紹以下三個和CWN有關的套件
  - CwnGraph： 用Python API取用CWN資料
  - CwnSenseTagger: CWN的詞意消岐模型
  - [optional] DistilTag: 斷詞及POS標記
    - 同樣可以使用CkipTagger
- 範例程式請見colab notebook / nbviewer

環境安裝

- colab環境下，套件和資料/模型都可順利安裝。
- 在本機環境下，請留意以下幾點:
  - 如果您之前有從GitHub上直接clone過CwnGraph，請留意版本是不同的。請您重新pull一次，或直接使用PyPI的最新版本。
  - CwnSenseTagger, DistilTag都需要pyTorch 1.6及transformers 3.2，請留意版本或另外建立虛擬環境。
  - 在Windows環境下安裝pyTorch請參考官方網站的指引。

CwnGraph: CwnBase

- CwnGraph 的目的是讓程式開發者能方便地用Python API取用、操作CWN的資料。




```
str(sense_tree), sense_tree.definition, sense_tree.all_examples(), sense_tree.semantic_relations
```

```
('<CwnSense[05235601](樹): 木本植物類的通稱，多年生植物，主要由樹幹及樹葉組成。>',
 '木本植物類的通稱，多年生植物，主要由樹幹及樹葉組成。',
 ['昨天晚上吹了一陣西風，把一棵碧綠的<樹>凋零了。',
  '這些<樹>不就是因為沒有用，所以才能長得這麼高大？',
  '因為當地人士要闢建停車場，所以要將預定地上的一棵二百年老<樹>移開。',
  '這後山的<樹>，因為不成材而得以逃過人們砍伐之劫，才有這般的悠然自在。'],
 [('synonym', <CwnSense[09311801](樹木): 木本植物類的通稱，多年生植物，主要由樹幹及樹葉組成。>, 'forward'),
  ('synonym', <CwnSense[07034901](木): 木本植物類的通稱，多年生植物，主要由樹幹及樹葉組成。>, 'forward'),
  ('meronym', <CwnSense[05234201](樹皮): 樹的表皮。>, 'forward'),
  ('hypernym', <CwnSense[06658201](植物): 可自行製造養分，沒有神經、感覺且不能運動的生物。>, 'forward'),
  ('holonym', <CwnSense[08041501](樹林): 在同一區域生長的眾多樹木的集合體。>, 'forward'),
```

CwnGraph: CwnBase

- 請確定已經執行過CwnGraph.download(), 他會自動下載CWN資料。
- CwnBase是主要程式進入點, 它負責所有與CWN的互動。

```
import CwnGraph
CwnGraph.download()
from CwnGraph import CwnBase
cwn = CwnBase()
```

CwnGraph: 搜尋詞條

- 在CWN架構設計下，詞形（word forms）會對應多個詞條（lemmas），一個詞條會有多個詞意（senses）。所以，我們可以先從找詞條開始：

```
# lemmas is a List[CwnLemma]
lemmas = cwn.find_lemma("<regexpr here>")
# .senses is a List[CwnSense]
lemmas[0].senses
```

CwnGraph: 詞意資料與語意關係

- 每個詞意都是一個CwnSense物件，詞意的資料都可透過相關的屬性或方法取得，如定義、詞性、例句、語意關係。

```
sense_x.definition
sense_x.pos
sense_x.all_examples() # sense_x.example
sense_x.synonym
sense_x.semantic_relations
```

CwnGraph: 搜尋詞意資料

- 如果要搜尋所有的詞意資料，也可透過另外一個函式，它可用regex搜尋每個詞意對應的詞條、定義、和例句

```
cwn.find_senses(lemma="<regex>",
    definition="<regex>",
    examples="<regex>")
```

CWN資料 - CwnGraph

```
[4] cwn = CwnBase()
```

```
[5] cwn.find_lemma("^樹$")
```

```
[<CwnLemma: 樹_1>, <CwnLemma: 樹_2>]
```

To Jupyter Notebook

```
sense_tree = cwn.find_lemma("^樹$")[0].senses
sense_tree
```

```
[<CwnSense[05235601](樹): 木本植物類的通稱，多年生植物，主要由樹幹及樹葉組成。>,
 <CwnSense[05235602](樹): 比喻在商業上針對特定對象或管道經營。>,
 <CwnSense[05235603](樹): 形狀像樹的物體。>]
```

```
[7] sense_tree[0].synonym  # 同義詞
```

CwnSenseTagger: 從pipeline往下走

- 除了知道一個詞有哪些詞意，我們更想知道一個詞在句子裡是哪個詞意。
- CwnSenseTagger就是一個詞意消岐模型。
- 感謝辛苦的標記人員，幫助我們累積了上萬筆標記資料。加上原有的CWN例句，我們以Bert為基礎開發WSD模型。目前此模型正確率約8成左右。

斷詞與詞性標記

- CwnSenseTagger是站在巨人肩膀上往前走的一步：
  - CWN的「詞」和中研院平衡語料庫對「詞」的定義是一致的。
  - CWN的「詞性」也是使用詞庫小組(CKIP)的詞類標記集。
- 所以，CwnSenseTagger預設輸入資料是有CKIP詞類標註的。
- 坊間有許多斷詞和詞類標記系統可選擇，但大部分不是採用CKIP的詞類標記集。
- 目前中文斷詞和詞類標記模型正確率最佳的是CkipTagger。
- DistilTag也是訓練於平衡語料庫，是仍屬開發階段的模型。

CKIP POS Tag Set link

標籤	詞類	標籤	詞類
Na	一般名詞（如歌迷、舞台）	VA	動作不及物動詞（他在「唱歌」）
Nb	專有名稱（如人名）	VC	動作及物動詞（「表達」敬意）
Nep	指代定詞（如這、那）	VH	狀態不及物動詞（品質「優良」）
Neu	數詞定詞（如「二」個）	A	非謂形容詞（「共同」市場）
Nf	量詞（如一「條」）	Caa	對等連接詞（如與、和、及）
Nv	名物化動詞（如「主辦」單位）	Dfa	動詞前程度副詞（「最」近、很）

模型環境設置

- DistilTag和CwnSenseTagger都可用download()下載相關模型資料。
- DistilTag下載檔案大小約501M，CwnSenseTagger下載檔案約387M。

```
# pip install -U DistilTag CwnSenseTagger
import DistilTag
import CwnSenseTagger
DistilTag.download()
CwnSenseTagger.download()
```

DistilTag

- DistilTag會將輸入的字串斷句。輸出結果是一組句子，每個句子中包含斷詞和詞類標記結果。

```
tagger = DistilTag()
tagged = tagger.tag("<raw text>")
```

```
[[('他', 'Nh'),
  ('由', 'P'),
  ('昏沉', 'VH'),
  ('的', 'DE'),
  ('睡夢', 'Na'),
  ('中', 'Ng'),
  ('醒來', 'VH'),
  ('，', 'COMMACATEGORY')],
```

CwnSenseTagger: senseTag

- CwnSenseTagger的輸入格式正如DistilTag的輸出格式；其輸出結果就加在每個token後：List[List[Tuple[Word, POS, SenseID, SenseDef]]]

```
sense_tagged = senseTag(tagged)
```

```
[[['他', 'Nh', '05238501', '代指自己和對方以外的第三人。'],
  ['由', 'P', '04011906', '引介起始狀態。'],
  ['昏沉', 'VH', '', ''],
  ['的', 'DE', '09002801', '表以前述動作的狀態。'],
  ['睡夢', 'Na', '', ''],
  ['中', 'Ng', '04004618', '在事件的過程中。'],
  ['醒來', 'VH', '', ''],
  ['，', 'COMMACATEGORY', '', '']],
```

```
for sent in sense_tagged:
 for token in sent:
   sense_x = CwnSense(token[2], cwn)
   print("\t".join(token), end='\t')
   if sense_x.synonym:
     print(sense_x.synonym[0])
   else:
     print()
 print("----")
```

To Jupyter Notebook

```
他      Nh      05238501        代指自己和對方以外的第三人。
由      P       04011906        引介起始狀態。    <CwnSense[04014009](從): 引介
昏沉    VH
的      DE      09002801        表以前述動作的狀態。      <CwnSense[05225501](
睡夢    Na
中      Ng      04004618        在事件的過程中。  <CwnSense[04084907](中間): 在
醒來    VH
```

Thanks & Join us

吳由由、陳柏文、江琼玉、張鈺琳、
詞義標記圖隊

科技部 AI 創新研究中心專案【建構概念為本且具語義結合性的中文知識庫】(PI：馬偉雲、張俊盛、謝舒凱、許永真、陳克健)

最後問題與討論

Feedback & QA

Making sense of AI forum : how can WSD be in action