

word2vec 和 semantics

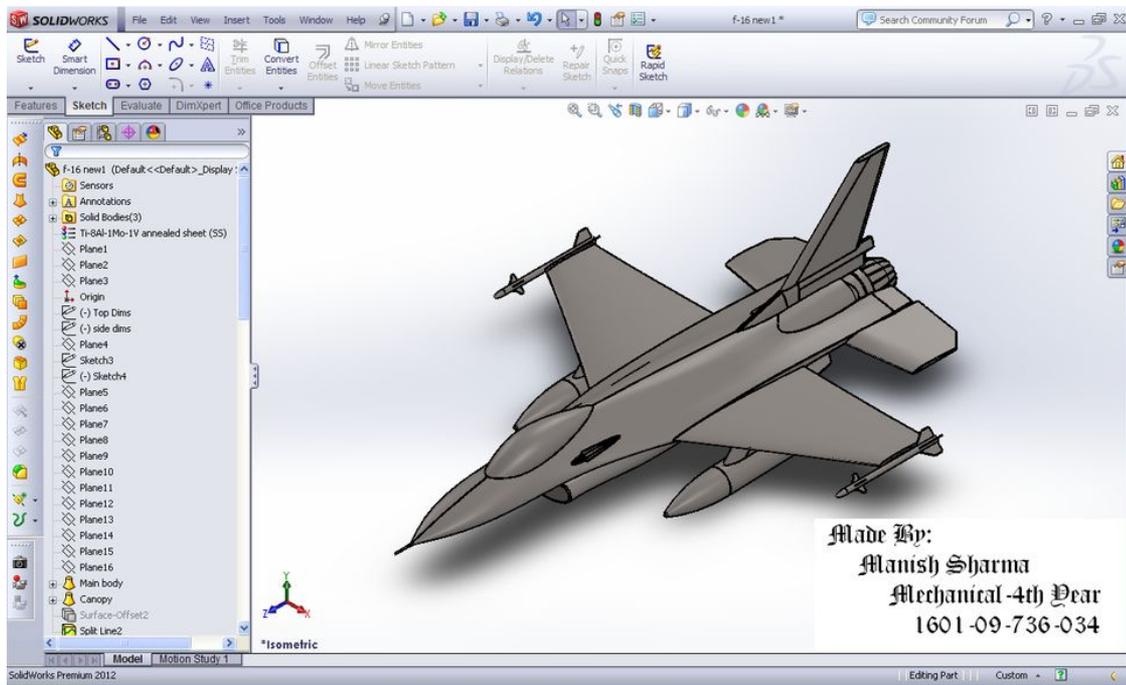
詞向量 與 語意

There's no magic here!

-PeterWolf

什麼是「模型」？

- 一組具體而「微」且可呈現原物特徵的記錄或實體



什麼是「語言模型」？

- 一組具體而「微」且可呈現**語言特徵**記錄或實體

Corpus Linguistics (語料庫語言學家):

簡單啊！我從語言裡做抽樣不就好了！

Some Computational Linguistics (計算機語言學家):

抽樣的程式我幫你寫，抽出來後的計算程式，我也可以幫你寫。

- 抽樣的前提：你知道母體長什麼樣子！
 - Question: 語言的母體(整體)長什麼樣子？你見過嗎？

NLTK: natural language toolkit

```
# 引用詞幹提取器
from nltk.stem import PorterStemmer
from nltk.stem import LancasterStemmer
from nltk.stem import SnowballStemmer

# 初始化
pst = PorterStemmer()
lst = LancasterStemmer()
snow = SnowballStemmer('english') //需定義語言

# 使用 (以Porter Stemmer為例)
pst.stem('eating')
pst.stem('eats')

# Result
eat
eat
```

```
>>> from nltk.stem import PorterStemmer
>>> pt = PorterStemmer()
>>> pt.stem("was")
'wa'
>>>
```

src. <https://medium.com/pyladies-taiwan/nltk-初學指南-三-基於-wordnet-的語義關係表示法-上下位詞結構篇-4874fb9b167a>

ref. [https://www.google.com/search?q="NLTK"+自然語言處理+課程](https://www.google.com/search?q=)

ref. [https://www.google.com/search?q="NLTK"+自然語言處理+選課](https://www.google.com/search?q=)

近代語言學史

理性主義 / 形式主義

語言學是獨立的學科



索緒爾

傳統語言學

人類學

語言教學

歷史語言學

(主要是音韻研究)

布倫菲爾德學派

語言是刺激與反射

哥本哈根學派

語言是代數

布拉格學派

語言是和社會現象
有關係的符號系統

結構主義



Chomsky

語言學是理論科學
(而且很像數學)

生成語法學派

語言可像數學一樣被計算，像科學一樣可證偽
橫空出世

現代語言學



放棄語言學

改宗 LLM

計算機語言學

語料庫語言學

(認知)功能學派

論文裡一定要提到 Chomsky 是錯的，語言是用來溝通的。
之後寫什麼無所謂。

行為主義 / 連結主義 / 反形式主義

Sparse vs. Continuous

離散 vs. 連續

蘿蔔

蘿 蔔的蘿是蘿蔔的第一個字而它的第二個字是 蔔

0.0 1.0

0.0 0.5 1.0

0.0 0.25 0.5 1.0

樹直徑 Tree Diameter

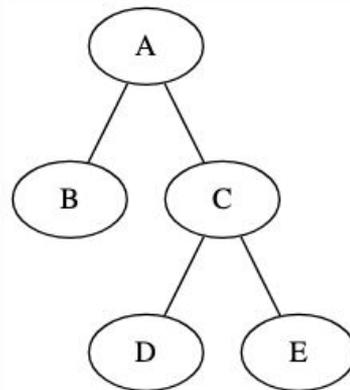
0

樹直徑定義為一棵 tree 上任兩點路徑中最大者，定義 $path(i, j)$ 為 vertex i 與 j 的路徑長度，則對所有 $1 \leq i \leq V$ 和 $1 \leq j \leq V$ 取 $\max(path(i, j))$ 即為樹直徑。

解法之一：分治

tree 上的路徑可由 root 分成兩種：經過 root 或者不經過 root。

經過 root 的路徑，兩端各在不同子樹，路徑長為 root 到兩端的距離總和，而任意點到 root 的距離即為其在樹上的深度，因此取子樹中深度最大的兩個，即為經過 root 的最長路徑。

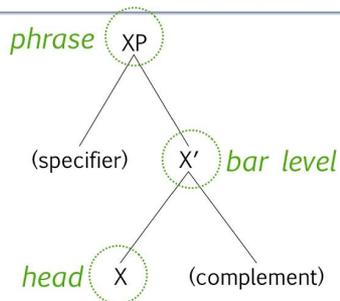


如上圖， E 到 root 的距離為 2，即 E 在樹上的深度。

描述「離散系統」最好的方式是「樹」

BTW, do you see an X-bar here?

How a phrase is organized



Modeling Features (and pretending it's semantics)

One-Hot Encoding & Word2vector



CKIP Lab
中文詞知識庫小組

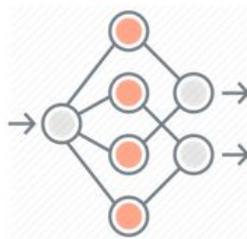
- 🏠 詞庫簡介
- 📄 最新消息
- 🔍 研究領域 ▾
- 📁 專案介紹 ▾
- 🖥️ 線上系統展示
- 📁 下載軟體與資源
- 📄 論文發表
- 👤 計畫主持人
- 👥 詞庫成員
- 📖 詞庫生活

CKIP Lab

中文 EN

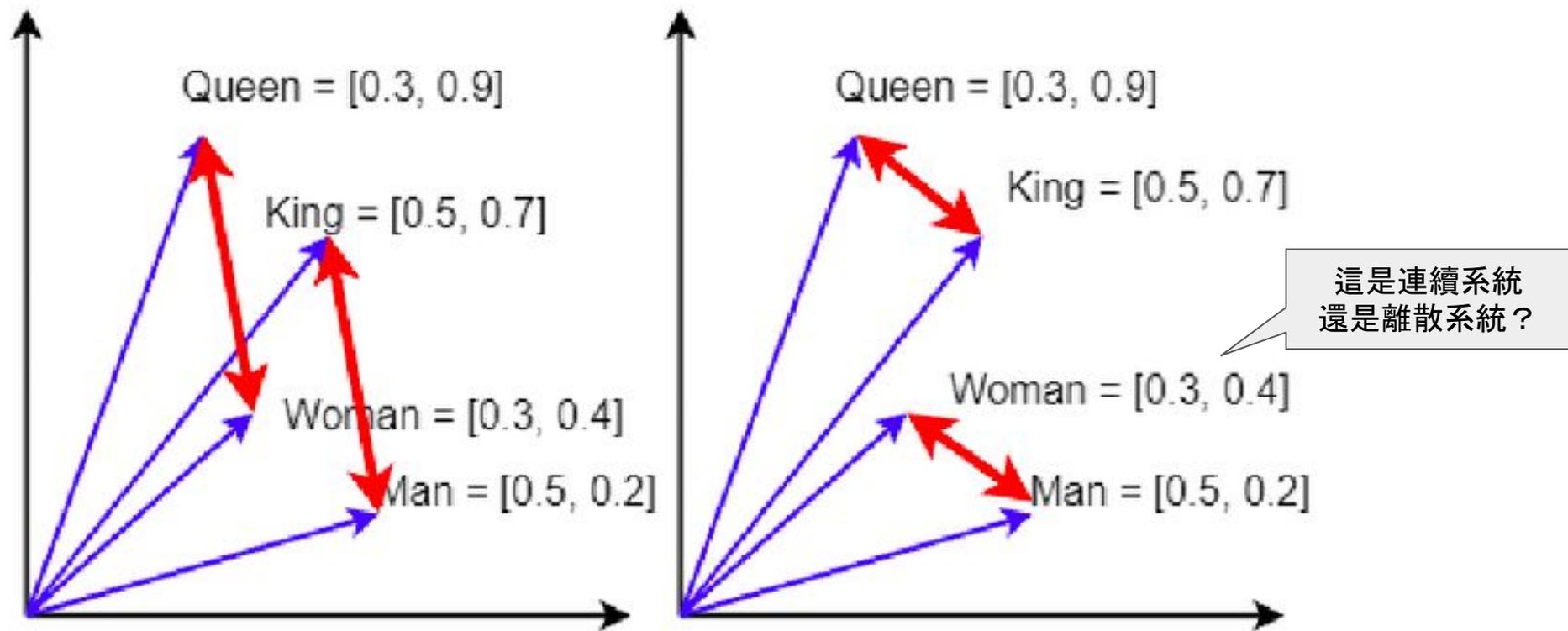
中文向量表達

「詞向量(word vector)」是將詞彙轉換成包含語意訊息的向量表達的技術。透過類神經網路訓練，將詞彙從 one-hot encoding 的高維度向量表達，轉換成低維度的向量，以利運算的進行。在向量空間中，cosine similarity 越高的兩個向量，也代表它們的語意越相近。



系統實作

1. 中文詞向量: 我們訓練中文詞向量的語料來源為 Chinese Gigaword 的中央社語料以及 ASBC 中研院平衡語料，共建構出 517015 詞彙訊息。目前提供兩個版本，建立方式與參數分別說明如下：
w2v : word2vec (skip-gram), 300 維, 其它參數為預設. (zip, 559MB)
Glove : glove, 300維, 其它參數為預設. (zip, 544MB)
2. 中文詞向量評估資料集：目前中文最大規模的詞向量評估資料集，建構細節請參以下論文: Chi-Yen Chen, Wei-Yun Ma. "Word Embedding Evaluation Datasets and Wikipedia Title Embedding for Chinese". LREC, May 2018.



The classical king + woman - man \approx queen example of neural word embeddings, in 2D. It must follow that king - man \approx queen - woman, and we can visually see that in the red arrows. There are 4 analogies one can construct, based on the parallel red arrows and their direction. This is slightly idealized; the vectors need not be so similar to be the most similar from all word vectors. The similar direction of the red arrows indicates similar relational meaning.

	A	B	C	D	E	F	G	H	I	J	K
1		1	2	3	4	5	6	7	8	9	10
2	(a)	我	其	實	並	不	認	識	這	個	人
3	(b)	我	其	實	並	不	認	識	那	個	人
4											

<https://docs.google.com/spreadsheets/d/14OJ3BcNeteWH40IMYjtiHsimGL5AkpwZBHJKyodzIE/edit?usp=sharing>

圖 src.

<https://blog.csdn.net/keeppractice/article/details/106177554>

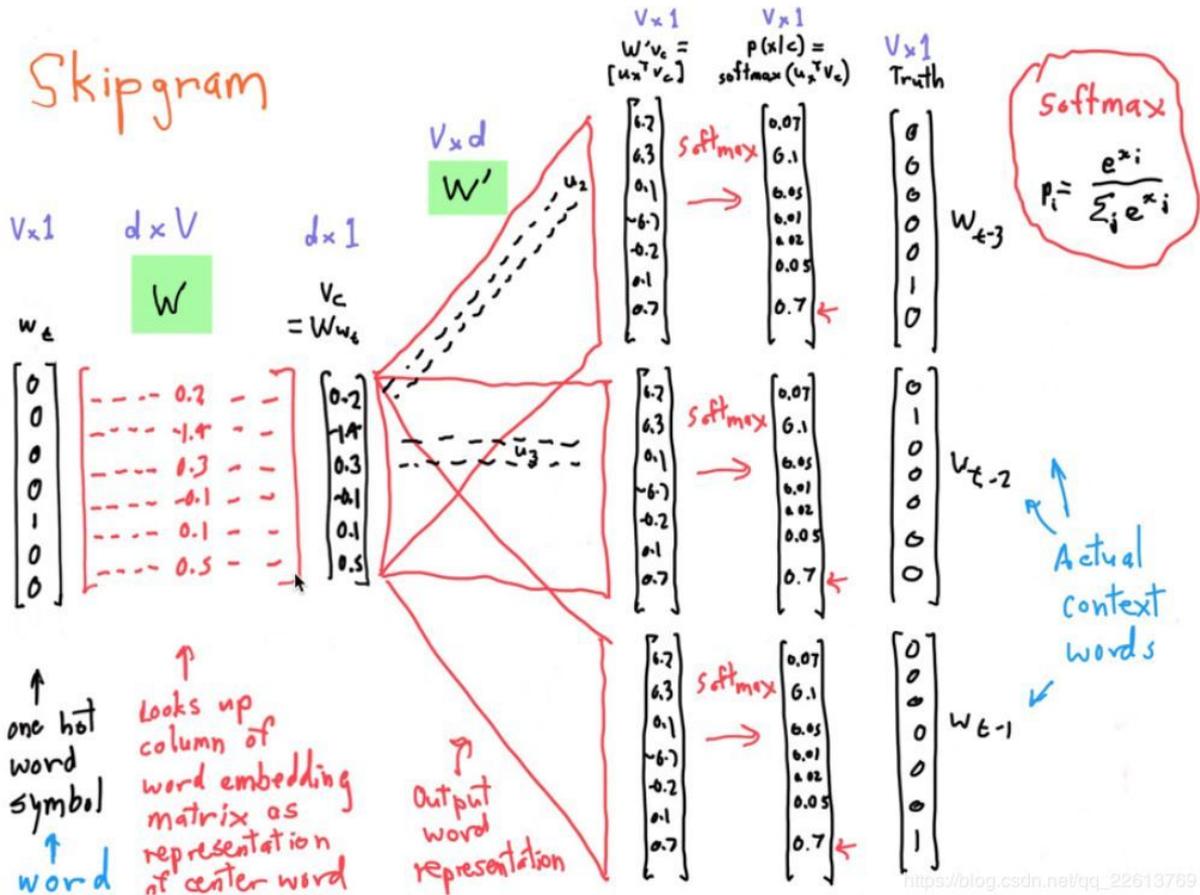
$$[0 \ 0 \ 0 \ 1 \ 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \ 12 \ 19]$$

你幹嘛把一個離散系統，
變成連續系統？

$$[0 \ 0 \ 0 \ 1 \ 0] \times \begin{bmatrix} 0.08 & 0.12 & 0.00 \\ 0.11 & 0.02 & 0.03 \\ 0.02 & 0.03 & 0.06 \\ 0.05 & 0.06 & 0.09 \\ 0.05 & 0.09 & 0.12 \end{bmatrix} = [0.05 \ 0.06 \ 0.09]$$

因為我不知道原來生成語言學已經發展出 X-bar Tree 了...我以為的 tree 還是 1956 年的那個...

Skipgram



如果有兩個字符，它們的前後總是差不多的字。那麼就表示，它們最後乘出來的輸出序列，在數值上會很接近！

即，在向量空間中，它們的位置會很接近！

親手玩玩看 Vector ! (至少玩兩次！)

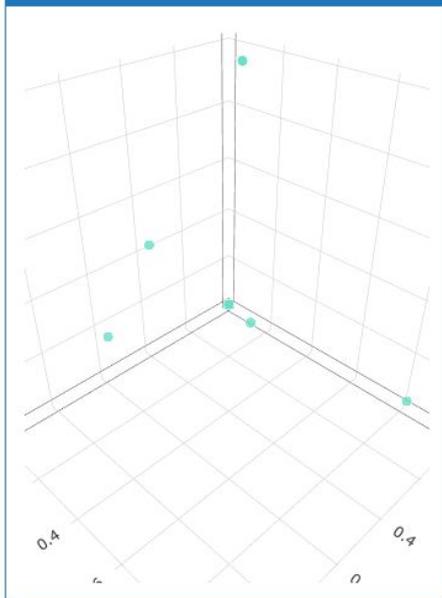
Training data (Corpus)

我其實並不認識這個人我其實並不認識那個人

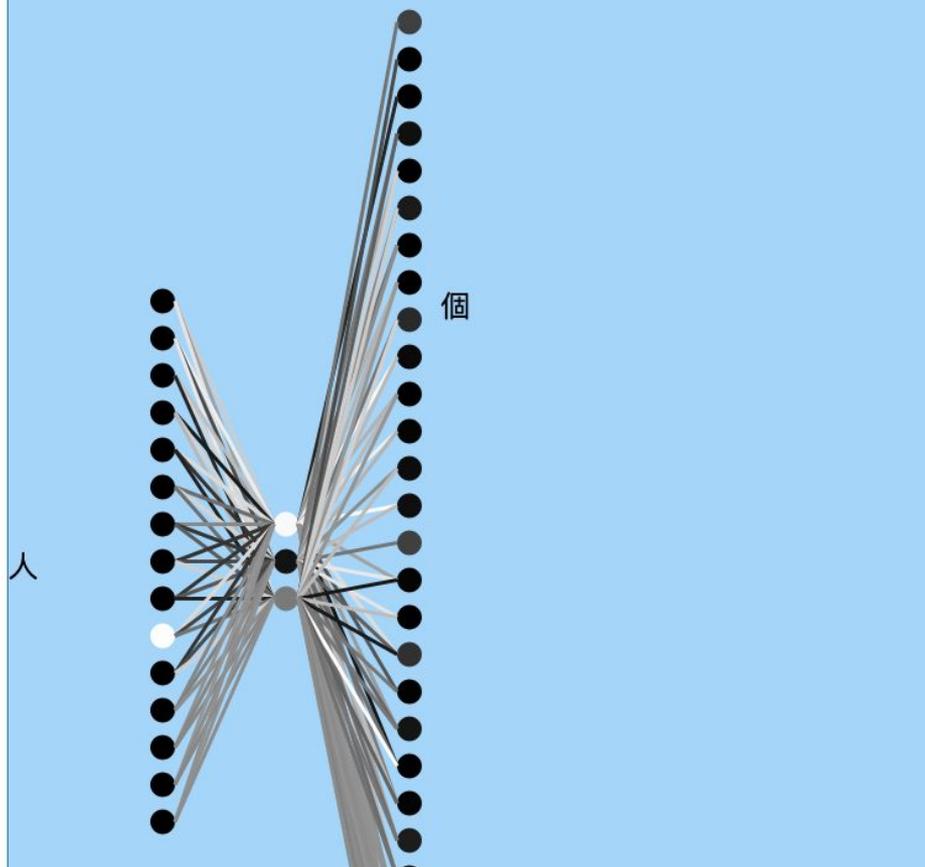
個人

Train

Vector space



Neural Network Visualization





限閱：p23, p24, p25

🕒 課程目標

1. Language as a human behavior: a complex adaptive usage-based system
2. Cognitive linguistics: Its origin and how it differs from formal, functional linguistics
3. Grammar and cognition
4. Word order and clause structure
5. Word classes and phrasal structure
6. Major clause types and multi-clausal constructions
7. Grammaticalization and the gradience of categories
8. Syntax/morphology interface:
Grammatical relations (subject and object)
9. Syntax/semantic interface: Semantic relations (agent and patient)
10. Construction Grammar vs. Cognitive Grammar
11. Conceptual metaphor and conceptual blending

一直忘記分享一件很震驚的事：
台大的句法學不學畫樹！！！！

以「向量」做為語意

- 企圖:用一個「字/字符」的前後環境, 做為這個「字/字符」的語意
- 實做:
 - 把語意利用 one-hot encoding 轉為數值 (因為這樣比較好算),
 - 再把數值壓到 0~1 之間, 成為高維向量 (因為這樣比較好算),
 - 再把高維向量分佈中「接近」的字字符, 視為語意接近。(因為看起來好像是這樣)
- 問題:
 - 前後文類似 (即高維空間中接近) 的字/字符, 真的有一樣或類似的語意嗎?



干安捏

Modeling Semantics

天真又業餘的語意學家

- Prototyping:
 - 人類在溝通的時候，傳達的語意，必然帶有某種「原型」。
 - 約翰對瑪麗說「草地上有一隻麻雀」！
 - 約翰和瑪麗心中必然出現「草地」的原型影像。
 - 約翰和瑪麗心中必然出現「麻雀」的原型影像。
 - 於是約翰就把「草地上有一隻麻雀」的語意，傳遞給瑪麗了！



語意在傳遞一個「真值條件」(或是數學上的函式啦)

- 約翰對瑪麗說「草地上有一隻麻雀」！
 - 此時發生的事情如下...
 - 約翰說的話，其語意被包裝成...
 - **在我的認知世界 w 中，有一個東西，它屬於「麻雀集合體」的一份子；還有一個東西，它屬於「草地集合體」的一份子。這個麻雀集合體的一份子，位置在草地集合體的一份子的上面。**
 - 上面這一段傳遞給瑪麗，請瑪麗檢查一下是不是一個「可解」的函式(真值條件)？
 - 瑪麗確認「是，這個函式可解，我收到它的真值條件了」。那麼瑪麗就聽懂了。

蝦毀？我不信！我的腦子真的這樣運作？



讓我馬上做一個對比實驗！

你聽得懂這一句話，所以你沒反應：

- 餐桌上有一台血壓機

你聽不懂這一句話，

你心裡會冒出一個「蛤？」

- 小明今天有要到海掃金

在我的認知世界 w 中，有一個東西 X ，它屬於「血壓機集合體」的一份子，還有一個東西 Y ，它屬於「餐桌集合體」的一份子 ...

在我的認知世界 w 中，有一個東西 X ，它屬於「海」的一份子？蛤？

讓我馬上做第二個對比實驗！

你聽得懂這一句話，所以你沒反應：

- 小明今天收到海掃金

你聽不懂這一句話，

你心裡會冒出一個「蛤？」

- 掃今海金收小明天到

在我的認知世界 w 中，有一個東西 X ，它屬於「海掃金集合體」的一份子

...

在我的認知世界 w 中...蛤？

實驗一告訴我們什麼？

語意的傳遞：只要你傳一個「可以解」的函式就行了。這個「可以解」的函式，在語意學的術語裡叫做「具有真值條件」。

[[餐桌上有一台血壓機]]

$\lambda x \lambda y. \text{ON}(\text{BP_MACHINE}(x), \text{TABLE}(y))^w$

[[小明今天有要到海掃金]]

$\lambda x \lambda y. \text{ACQUIRE}(\text{HEXOGEN}(x), \text{MING}(y))^w_{\text{today}}$

有沒有發現，你**不用真的認識小明**，你甚至**不用知道海掃金長什麼樣子**，但你心裡沒有「蛤」了！因為你「知道了這個函式是**可以解**的」。

實驗二告訴我們什麼？

[[到天海掃金今收小明]]

$\lambda x \lambda y. \text{HEXOGEN}(x), \text{MING}(y)$

你**不用真的認識小明**，你甚至**不用知道海掃金長什麼樣子**，

但是，你心裡還是有個「蛤」！因為你「知道這個函式**無法解**的」。

語意學的術語叫「它沒有真值條件」(因為它的句法樹是錯的)

人類怎麼理解「語意」的呢？只要有結構，就能計算語意！

How many red cars in Palo Alto does Kathy like?

$S' : | \lambda x. \text{car}(x) \wedge \text{in}'(\text{paloalto})(x) \wedge \text{red}'(x) \wedge \text{like}(x)(\text{kathy}) |$

$\text{NP}[\text{wh}] : \lambda V. | \lambda x. \text{car}(x) \wedge \text{in}'(\text{paloalto})(x) \wedge \text{red}'(x) \wedge V(x) |$

Aux

$S : \lambda z. \text{like}(z)(\text{kathy})$

$S/\text{NP} : \text{like}(z)(\text{kathy})$

$\text{Det} : \lambda P. \lambda V. | \lambda x. P(x) \wedge V(x) |$

does

$N' : \lambda x. \text{car}(x) \wedge \text{in}'(\text{paloalto})(x) \wedge \text{red}'(x)$

$\text{NP} : \text{kathy}$ $\text{VP}/\text{NP}_z : \text{like}(z)$

How many

$N' : \lambda x. (\text{car}(x) \wedge \text{red}'(x))$

$\text{PP} : \lambda P. \lambda x. (P(x) \wedge \text{in}'(\text{paloalto})(x))$

Kathy

$V : \text{like}$ $\text{NP}/\text{NP}_z : z$

$\text{Adj} : \lambda P. (\lambda x. P(x) \wedge \text{red}'(x))$

$N' : \text{car}$

$P : \lambda y. \lambda P. \lambda x. (P(x) \wedge \text{in}'(y)(x))$

$\text{NP} : \text{paloalto}$

like

e

red

N : car

in

Palo Alto

cars

```
select count(*) from Likes, Cars, Locations, Reds where Cars.obj =
Likes.liked AND Likes.liker = 'Kathy' AND Red.obj = Likes.liked AND
Locations.place = 'Palo Alto' AND Locations.obj = Likes.liked
```

Zenith: symbolic ML "semantic parsing": Zettlemoyer & Collins 2005; Artzi & Zettlemoyer 2013; Liang, Jordan & Klein 2013

Modeling Features

- Word2Vector:
 - 儲存每一個「字/字符」的前後關係
 - 關係需要「大量刺激」才能被穩定儲存
 - 反正是...「前後」，所以順序與結構不那麼重要

Modeling Semantics

- Semantics:
 - 計算每一個「實體 (Entity)」之間的關係
 - 關係是「現場計算」的，見過一次就行了 (所以不需要大數據)
 - 句法結構很重要！

Step hamburger me step, a by stack how teach to



思考 7 秒 ▾

Sure! Here's a step-by-step guide on how to stack a hamburger:

Ingredients:

- **Hamburger Buns:** 1 set (top and bottom)
- **Hamburger Patty:** 1 (beef, turkey, veggie, or your preferred type)
- **Cheese Slice (Optional):** 1 slice (cheddar, American, etc.)
- **Lettuce Leaf:** 1 large leaf or several smaller ones
- **Tomato Slices:** 2–3 slices
- **Onion Slices:** 2–3 rings (red or white onion)
- **Pickle Slices (Optional):** 2–3 slices
- **Condiments:** Ketchup, mustard, mayonnaise, or your favorites

<https://chatgpt.com/share/66e3ab51-b2a8-800e-8d5e-156018387dbd>

Equipment:



理論與公式

Word2Vec有兩種架構：連續詞袋（CBOW）和Skip-Gram。CBOW預測目標詞基於上下文詞，而Skip-Gram正好相反，它預測上下文詞基於目標詞。

Skip-Gram模型

Skip-Gram模型的目標是最大化給定目標詞的條件下上下文詞的出現機率，即最大化下面的目標函數：

幹嘛要用倒數？
這樣我可以確保最大不會超過 1

$$L = \frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

幹嘛加起來？
連乘的值，取 log 以後改為連加

幹嘛取 log？
因為小數連乘，愈來愈小，電腦不會算 ...

其中， T 是文字中的詞彙總數， c 是上下文大小， w_t 是目標詞， w_{t+j} 是上下文詞。 $p(w_{t+j} | w_t)$ 通常使用softmax函數來計算：

和前一個有什麼關係？
為了保有順序資訊。

$$p(w_O | w_I) = \frac{\exp(v'_{w_O} v_{w_I})}{\sum_{w=1}^W \exp(v'_w v_{w_I})}$$

取 $\exp()$ 又相除，在幹嘛？
為了把 $\{0, 1\}$ 變成 $0 \sim 1$ 的小數

這裡， v_{w_I} 和 v'_{w_O} 分別是輸入詞 w_I 和輸出詞 w_O 的向量表示， W 是詞彙表的大小。

Think again! Takeaway!

1. 語意是什麼？人類「理解」語言的面向。
2. 語言是離散系統，還是連續系統？
3. 向量是什麼？
4. 向量有沒有解決「理解語言」這個需求？所以「大型語言模型」懂不懂語意？
5. 人類怎麼「理解」語言的意義？
6. Prototyping (原型) / 圖像傳遞有用嗎？
7. 句法樹和語意的計算有沒有關係？
8. 句法樹是「離散系統」的描述方法，還是「連續系統」的描述方法？
9. 向量是「離散系統」的描述方法，還是「連續系統」的描述方法？
10. There is no magic!