

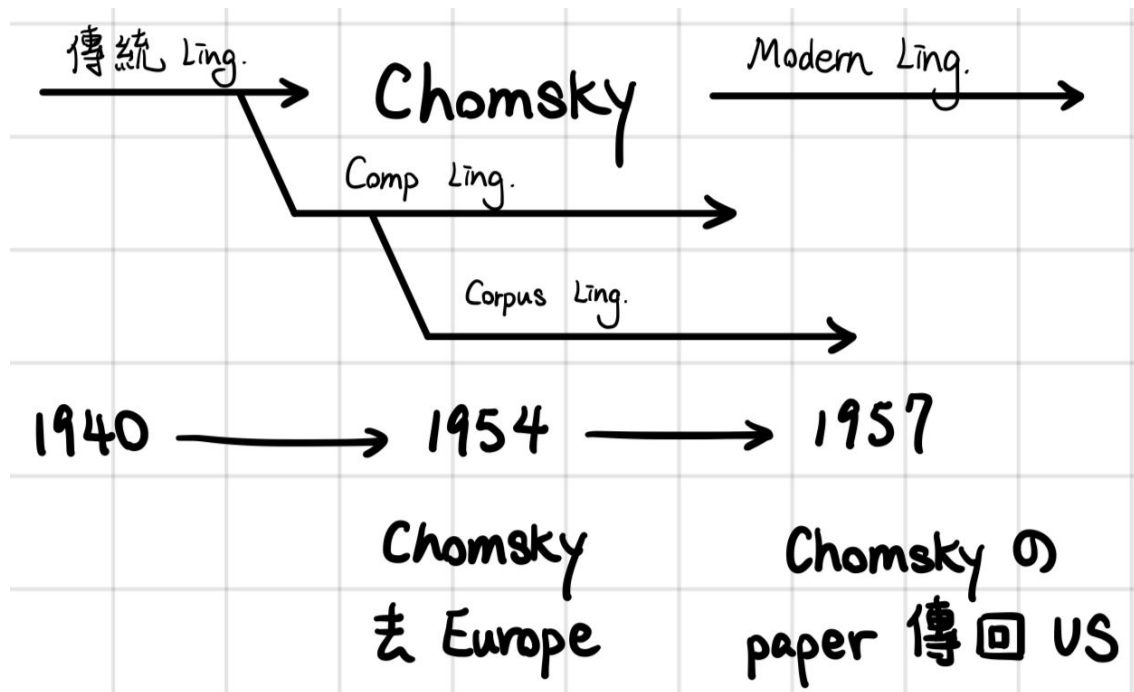
語意計算

(The Wrong Way)

一點點👉歷史

1. Saussure 把 Linguistics 從 Anthropology 分開

2.



Outline

Question: 怎麼知道兩個句子的意思像不像？

- I. Edit distance (編輯距離; Levenshtein Edit Distance)
- II. 座標系統上的兩點距離
- III. 高維座標系統的「維」
- IV. 向量是指什麼？
- V. 兩組向量的差異代表什麼？
- VI. 什麼是詞向量？

Edit Distance

table
↓
cable
comfortable

table → cable

t a b l e
↓
c a b l e

$$1+0+0+0+0=1$$

Algorithm	Operations Allowed			
	Insertions	Deletions	Substitutions	Transposition
Levenshtein Distance	✓	✓	✓	
Longest Common Subsequence (LCS)	✓	✓		
Hamming Distance			✓	
Damerau-Levenshtein Distance	✓	✓	✓	✓
Jaro distance				✓

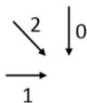
comfortable → table

c o m f o r t a b l e
↓ ↓ ↓ ↓ ↓ ↓
△ △ △ △ △ t a b l e

$$1+1+1+1+1+0+0+0+0+0=6$$

		M	O	N	K	E	Y
	0	1	2	3	4	5	6
M	1	0	1	2	3	4	5
O	2	1	0	1	2	3	4
N	3	2	1	0	1	2	3
E	4	3	2	1	1	1	2
Y	5	4	3	2	2	2	1

		M	O	N	K	E	Y
	0	1	1	1	1	1	1
M	0	2	1	1	1	1	1
O	0	0	2	1	1	1	1
N	0	0	0	2	1	1	1
E	0	0	0	0	2	2	1
Y	0	0	0	0	0	0	2

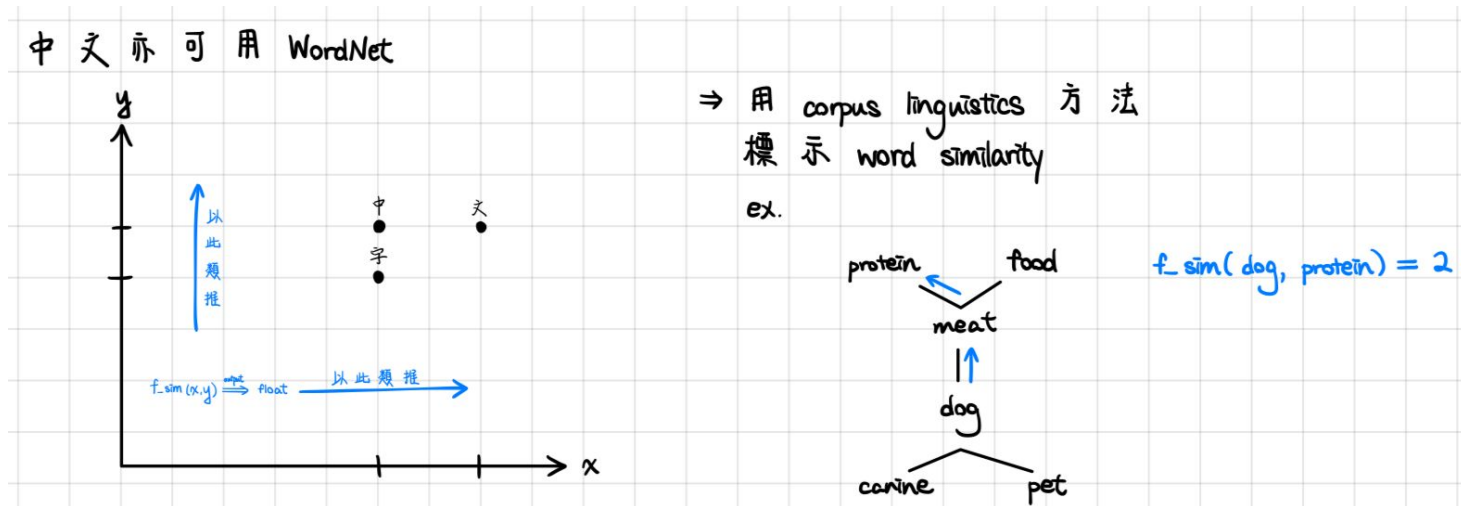


An example of the minimum edit distance algorithm. The left sub-figure show the F matrix, which is the minimum edit distance for the sub-sentences. The right sub-figure shows the path matrix, from which we can get the alignment.

Edit distance: 透過需要“edit”一個字而成為另個字的次數，去計算兩字之間的「距離」。

WordNet

另一個非字母形式的語言使用的方法: WordNet

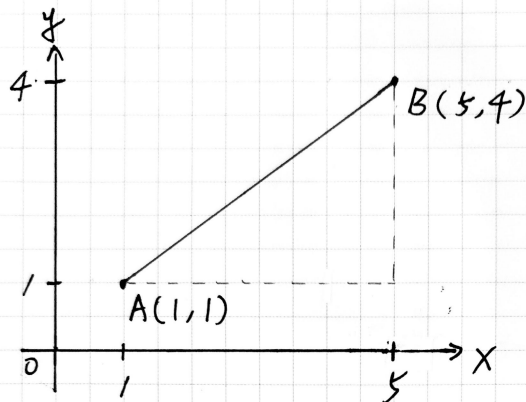


缺點: 人對於標註文字上的定義可能會意見分歧

座標系統上的兩點距離

座標系統上的兩個點可以透過畢氏定理取得兩點之間的距離。

假設有一個平面座標，由x軸與y軸構成，該座標上有A(1,1)、B(5,4)兩點，可透過計算得知兩點之間距離等於5。



$$\begin{aligned} \text{distance } \overline{AB} &= \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \\ &= \sqrt{(1 - 5)^2 + (1 - 4)^2} \\ &= 5 \end{aligned}$$

「峰」 versus 「峯」

雖然「峰」和「峯」意義上是一樣的，但是每個人的使用方法可能不一樣，例如人的姓名中，可能會使用其中一個字，且不會和另一個字在此情況下通用。

⇒ mapping 時 用 vectors
取 $\alpha(\text{中}, \text{字}), \beta(\text{字}, \text{文}), \theta(\text{中}, \text{文})$

