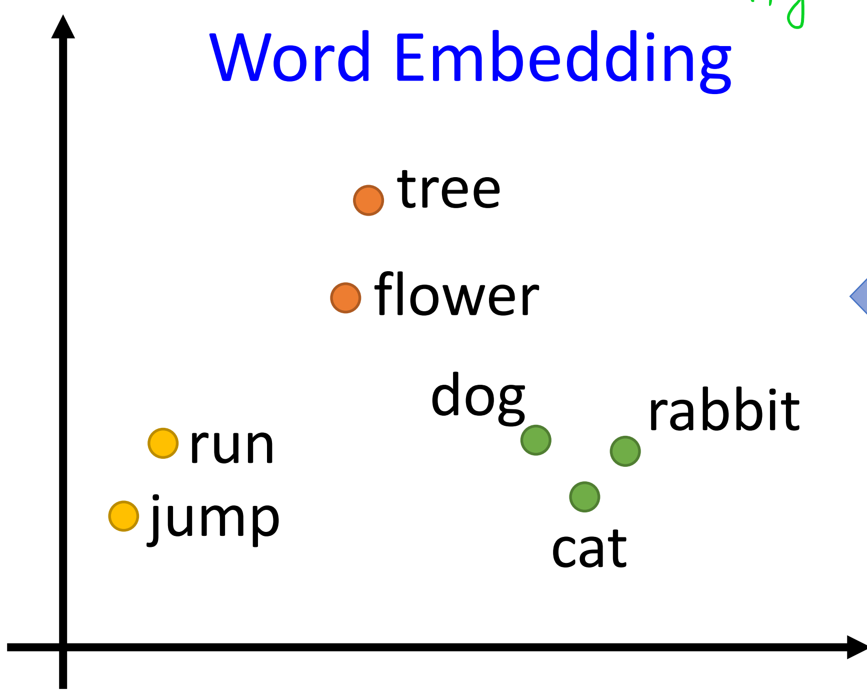
Unsupervised Learning: Word Embedding

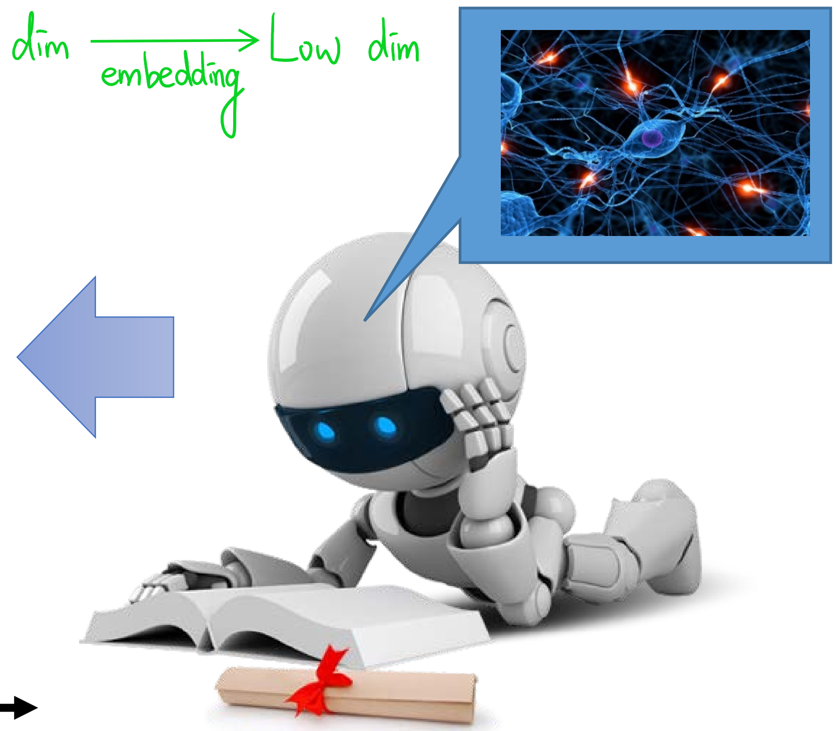
The dimension reduction on word.

Word Embedding

- Machine learns the meaning of words from reading a lot of documents without supervision

High dim $\xrightarrow{\text{embedding}}$ Low dim





Dim = # words

1-of-N Encoding

apple = [1 0 0 0 0]

bag = [0 1 0 0 0]

cat = [0 0 1 0 0]

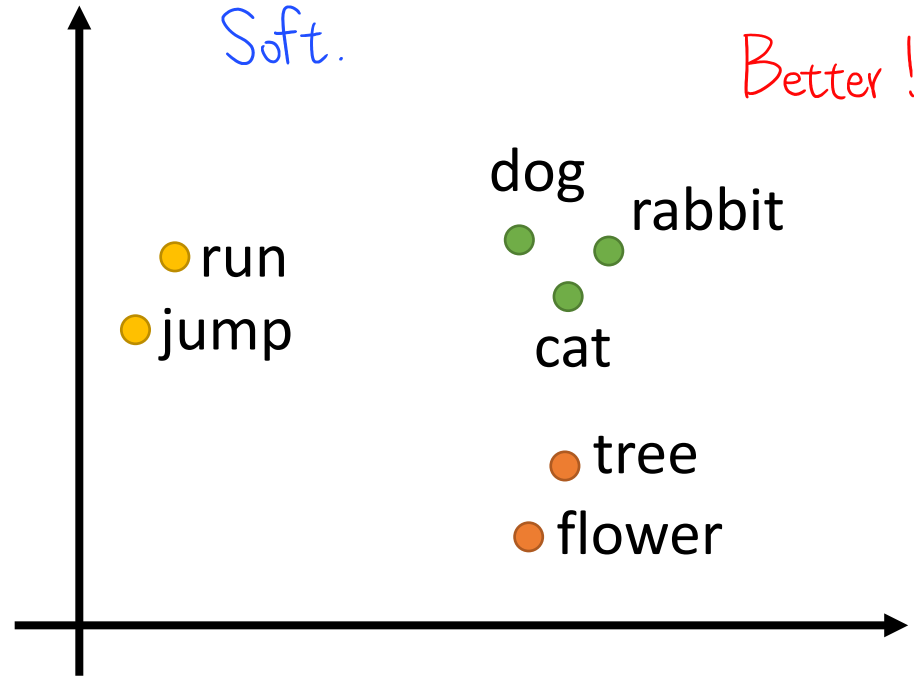
dog = [0 0 0 1 0]

elephant = [0 0 0 0 1]

Word Embedding

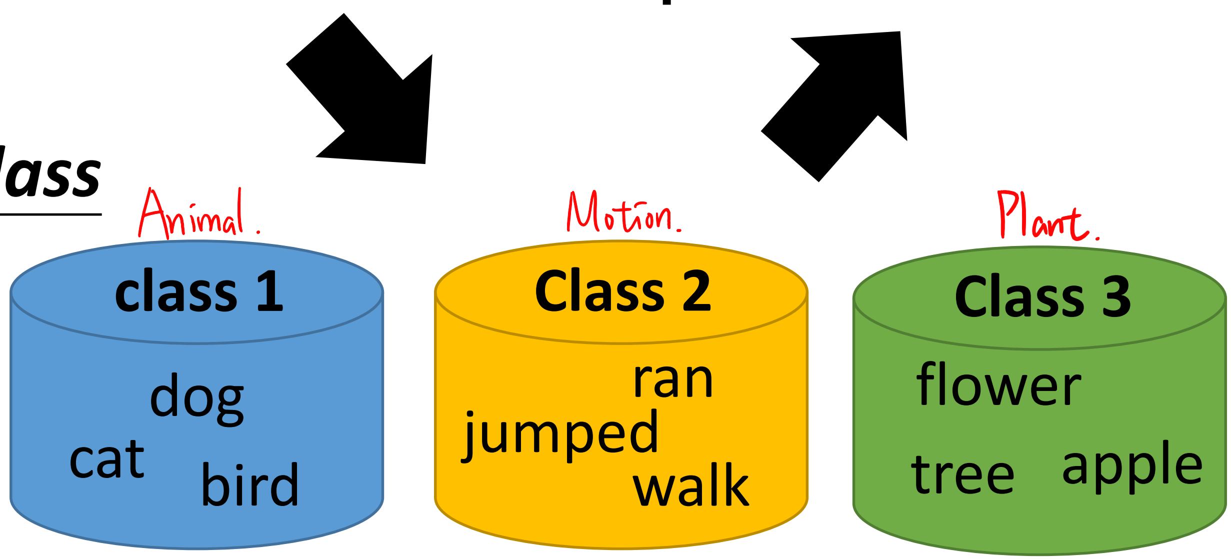
Soft.

Better !



Word Class

Hard.



Word Embedding

- Machine learns the meaning of words from reading a lot of documents without supervision
- A word can be understood by its context

You shall know a word by the company it keeps

蔡英文、馬英九 are something very similar

馬英九 520宣誓就職

蔡英文 520宣誓就職

How to exploit the context?

1° • **Count based** ex: P.22

- If two words w_i and w_j frequently co-occur, $V(w_i)$ and $V(w_j)$ would be close to each other
- E.g. Glove Vector: http://nlp.stanford.edu/projects/glove/

$$V(w_i) \,.\, V(w_j) \longleftrightarrow N_{i,j}$$

Positive correlation.

Inner product

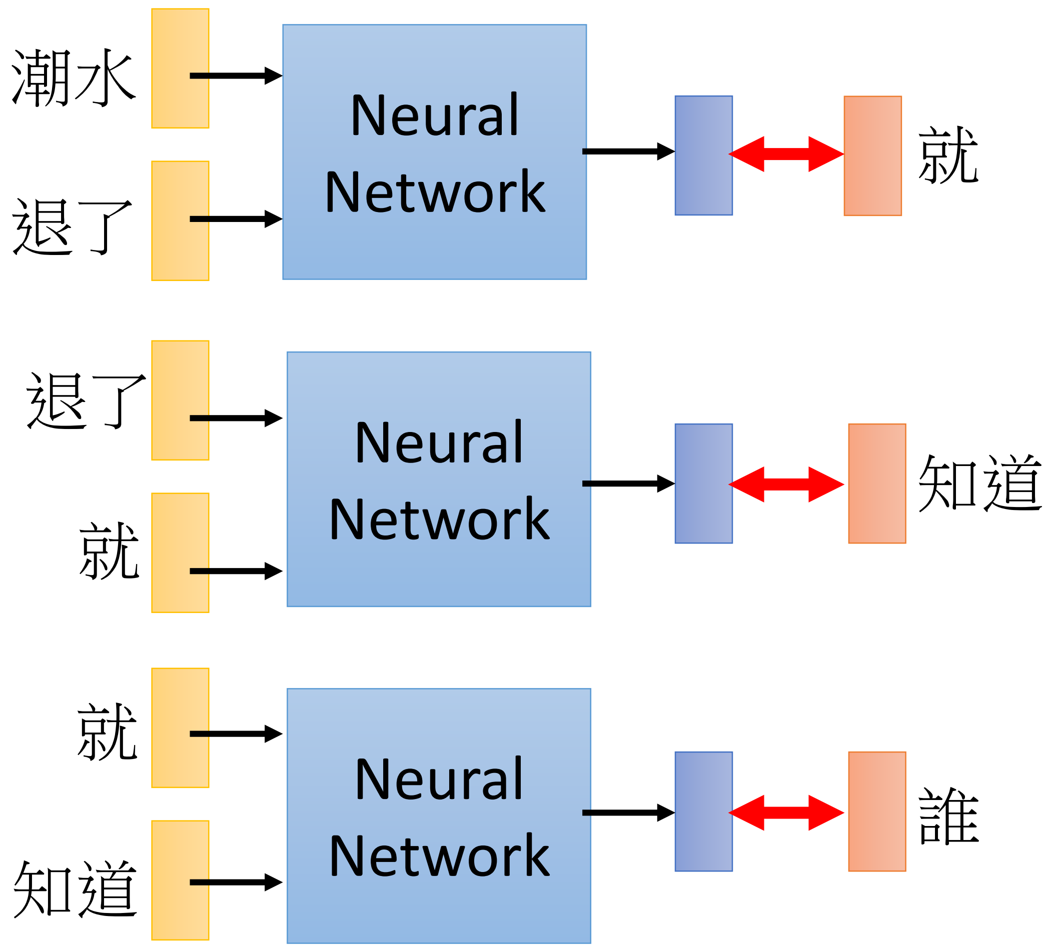
Number of times w_i and w_j in the same document

2° • **Predition based** Next page.

Prediction-based – Training

Collect data:

潮水 退了 就 知道 誰 …
不爽 不要 買 …
公道價 八萬 一 …
………



Minimizing cross entropy

Prediction-based - 推文接話

推 louisee :話說十幾年前我念公立國中時,老師也曾做過這種事,但
推 pttnowash :後來老師被我們出草了
→ louisee :沒有送這麼多次,而且老師沒發通知單。另外，家長送
→ pttnowash :老師上彩虹橋 血祭祖靈

https://www.ptt.cc/bbs/Teacher/M.1317226791.A.558.html

推 AO56789: 我同學才扯好不好，他有一次要交家政料理報告
→ AO56789:其中一個是要寫一樣水煮料理的食譜，他居然給我寫
→ linger:溫水煮青蛙
→ AO56789:溫水煮青蛙，還附上完整實驗步驟，老師直接給他打0
→ linger:幹還真的是溫水煮青蛙

著名簽名檔 (出處不詳)

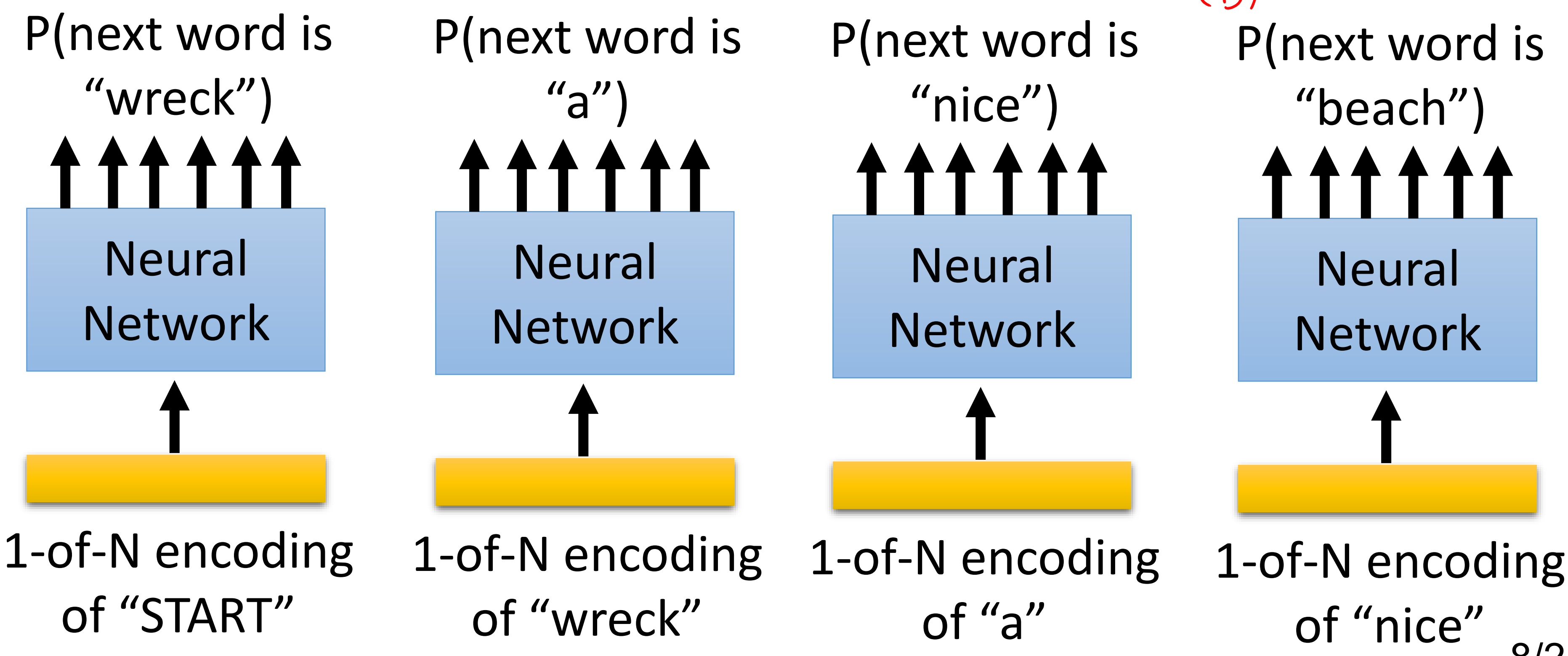
Prediction-based – Language Modeling

P(“wreck a nice beach”)

=P(wreck|START)P(a|wreck)P(nice|a)P(beach|nice)

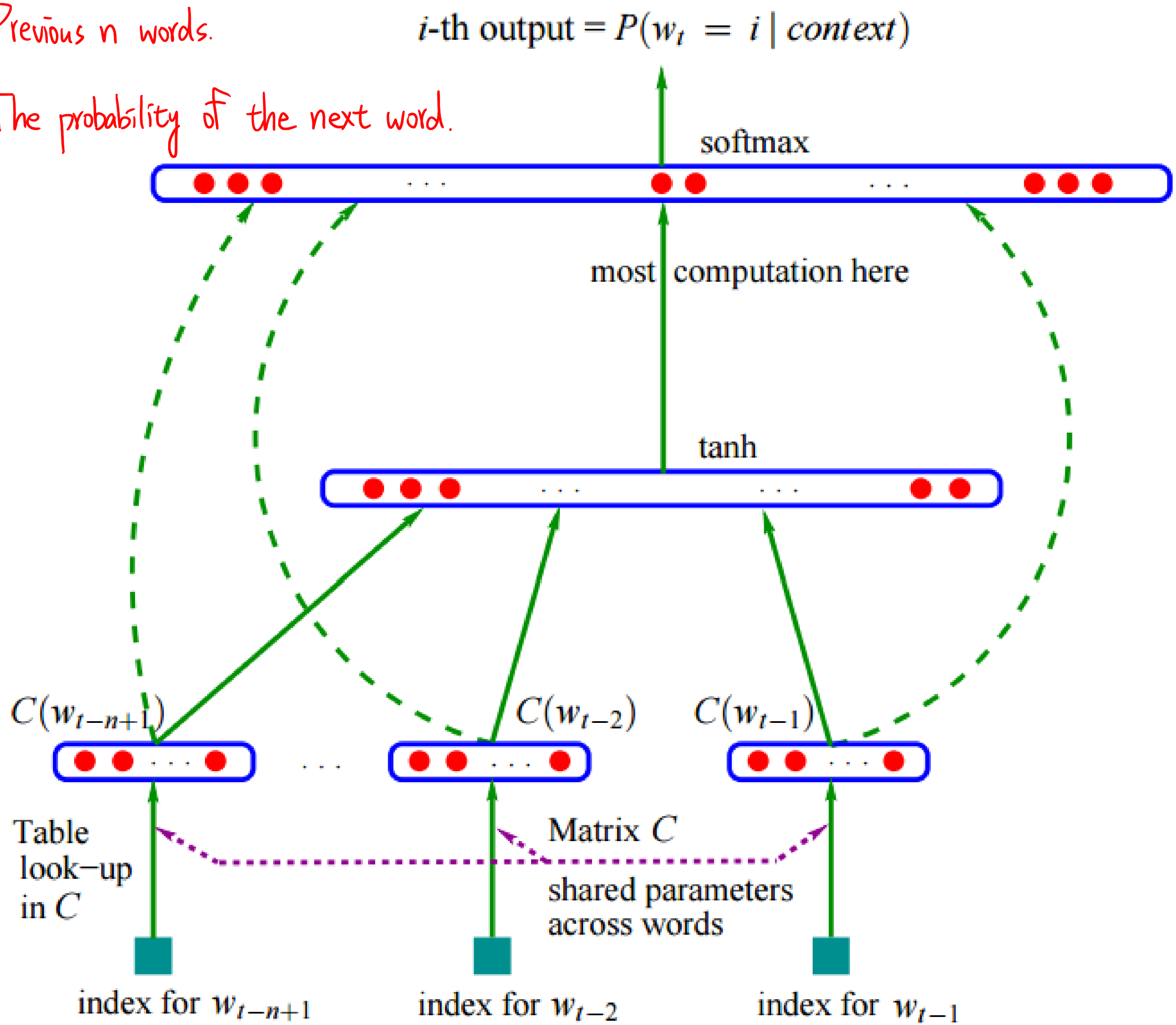
P(b|a): the probability of NN predicting the next word.

(b)



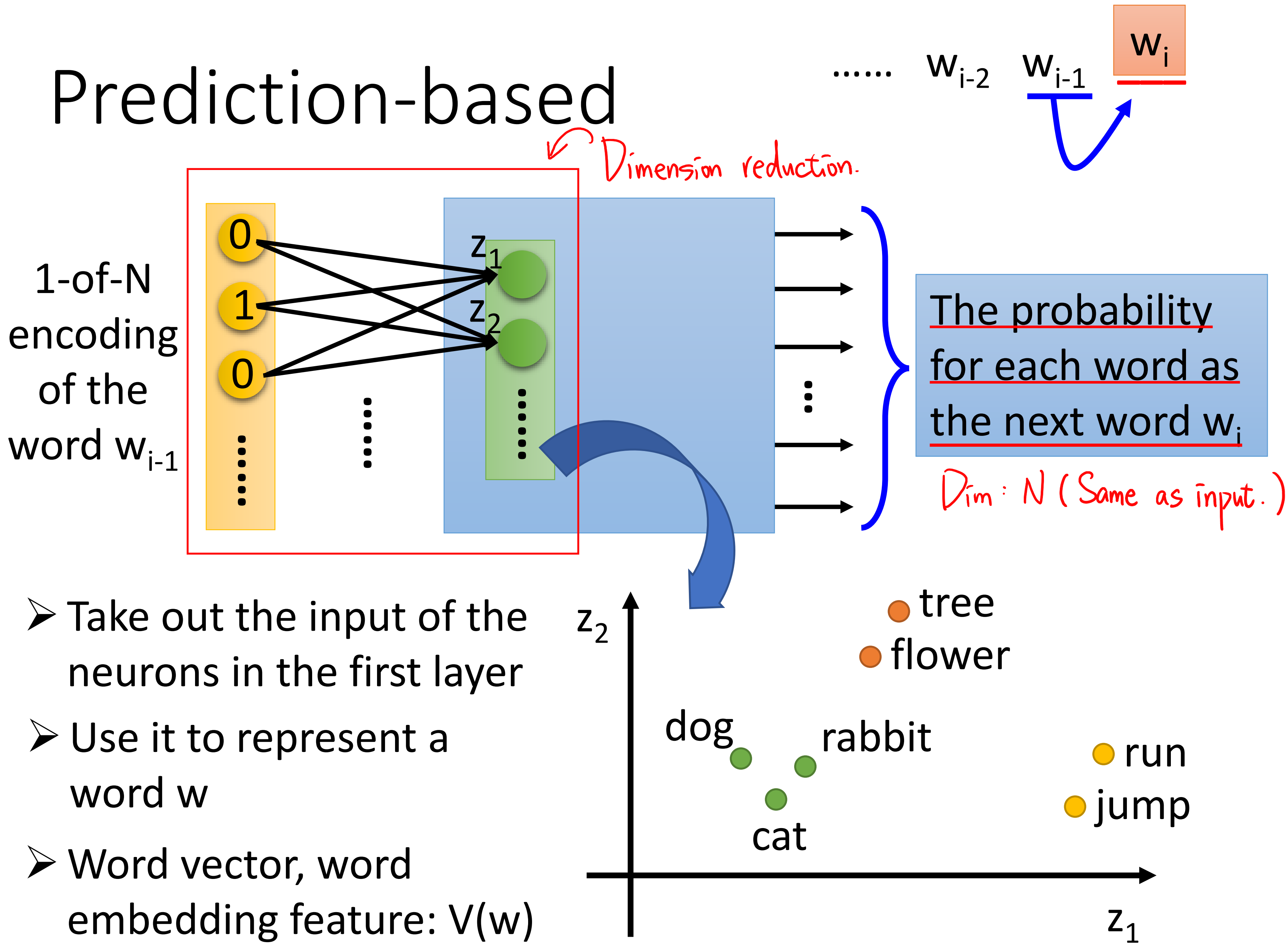
Input: Previous n words.

Output: The probability of the next word.



Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155.

Prediction-based



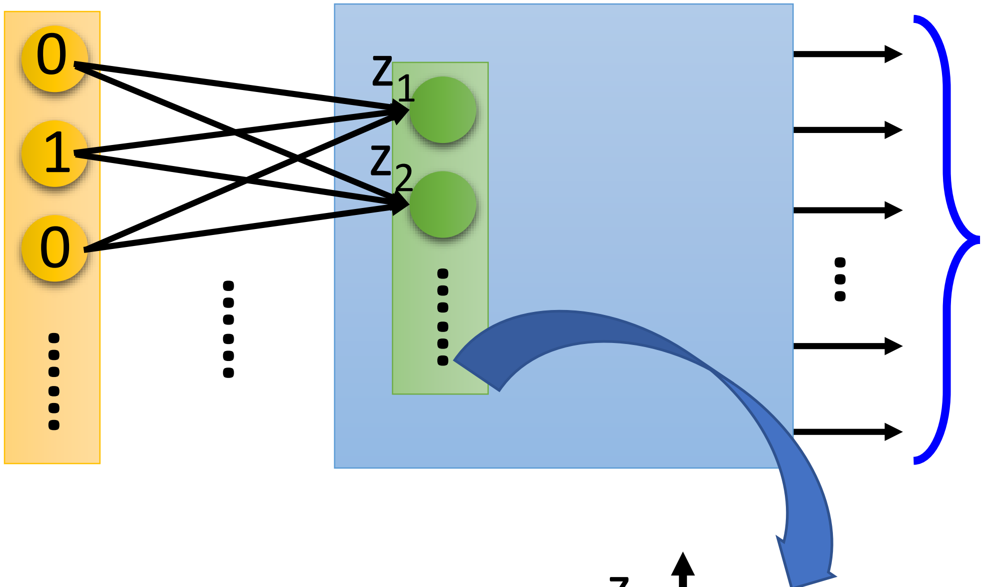
- Take out the input of the neurons in the first layer
- Use it to represent a word w
- Word vector, word embedding feature: V(w)

Prediction-based

But, You shall know a word by the company it keeps

Solution on the next page.

蔡英文
or
馬英九



The probability for each word as the next word w_i

"宣誓就職" should have large probability

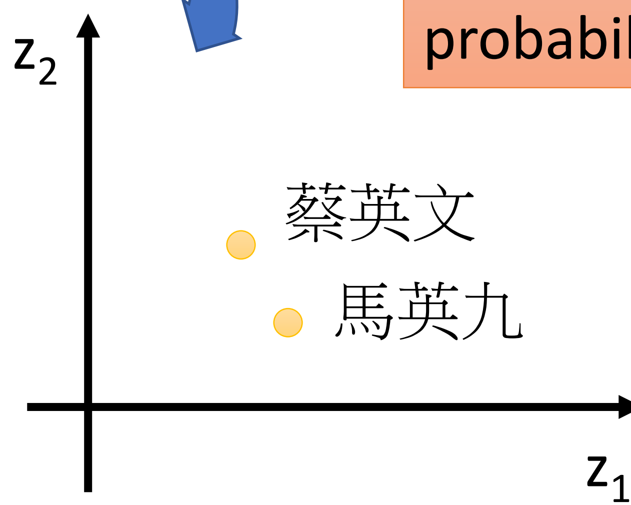
Training text:

…… 蔡英文 宣誓就職 ……

w_{i-1} w_i

…… 馬英九 宣誓就職 ……

w_{i-1} w_i

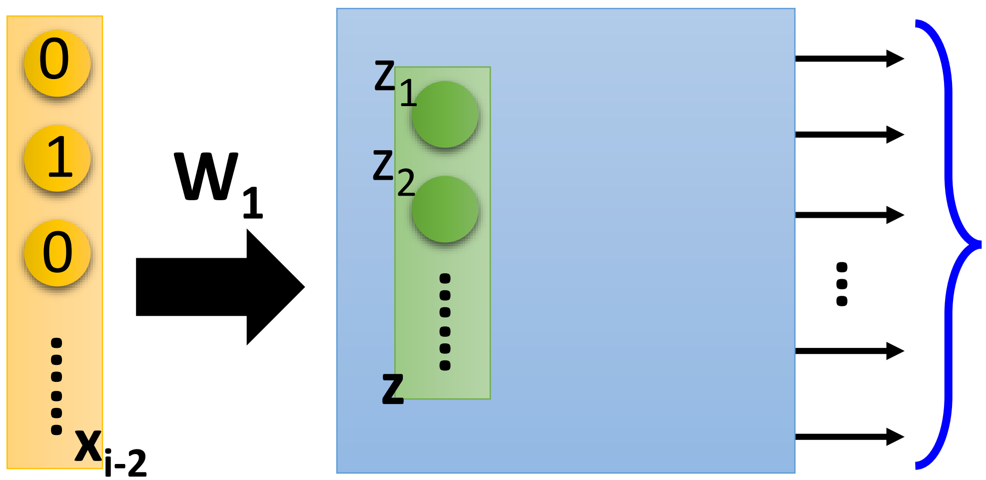


Prediction-based – Sharing Parameters

We want to see more previous words.

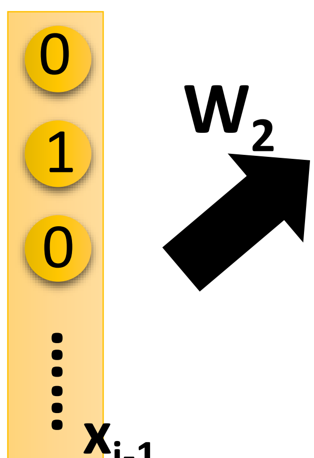
Problem.

1-of-N encoding of the word w_{i-2}



The probability for each word as the next word w_i

1-of-N encoding of the word w_{i-1}



The length of $\mathbf{x}_{i-1}$ and $\mathbf{x}_{i-2}$ are both |V|.

The length of $\mathbf{z}$ is |Z|.

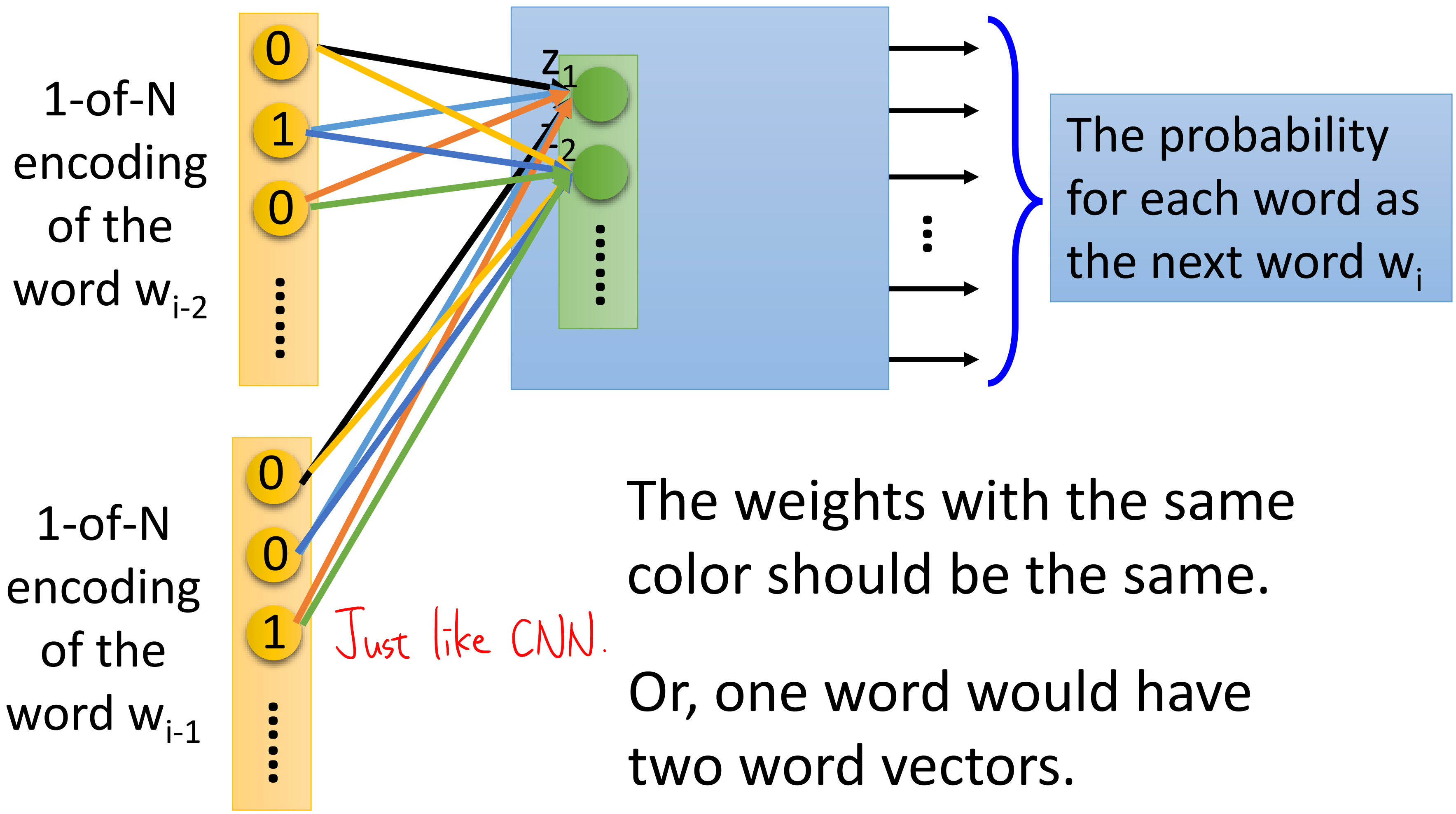
$\mathbf{z} = \mathbf{W}_1 \mathbf{x}_{i-2} + \mathbf{W}_2 \mathbf{x}_{i-1}$ Too many parameters.

The weight matrix $\mathbf{W}_1$ and $\mathbf{W}_2$ are both |Z|X|V| matrices.

Share W.

$\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W} \Rightarrow \mathbf{z} = \mathbf{W}(\mathbf{x}_{i-2} + \mathbf{x}_{i-1})$

Prediction-based – Sharing Parameters



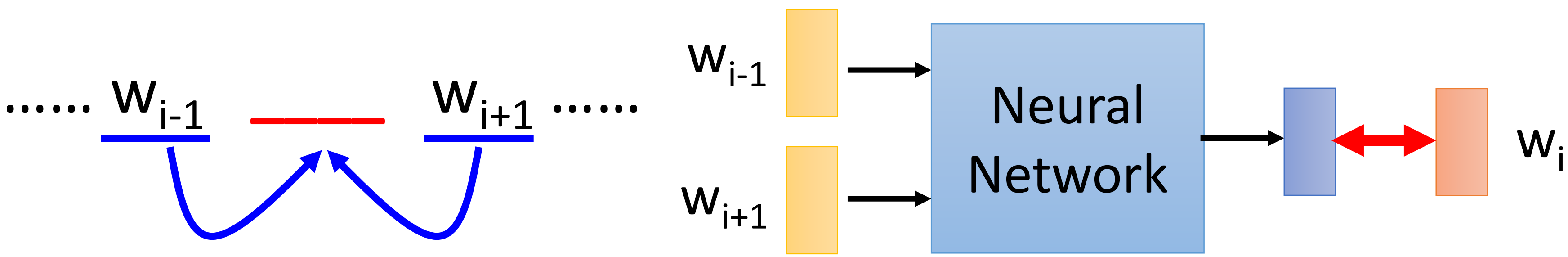
Just like CNN.

The weights with the same color should be the same.

Or, one word would have two word vectors.

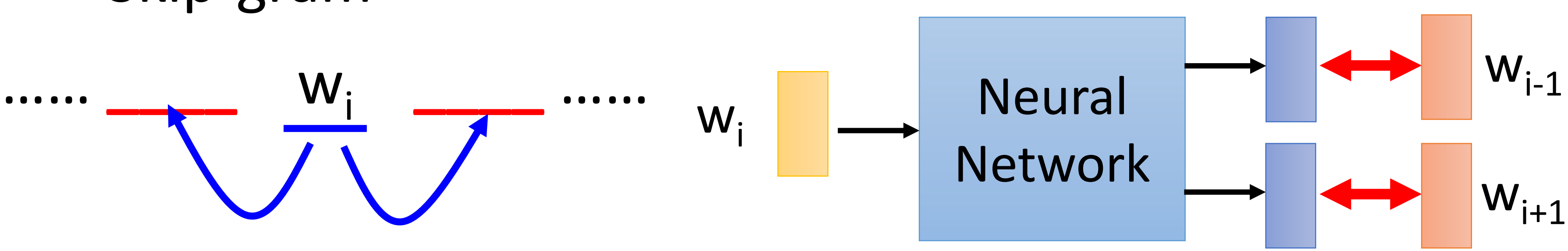
Prediction-based
– Various Architectures

- Continuous bag of word (CBOW) model



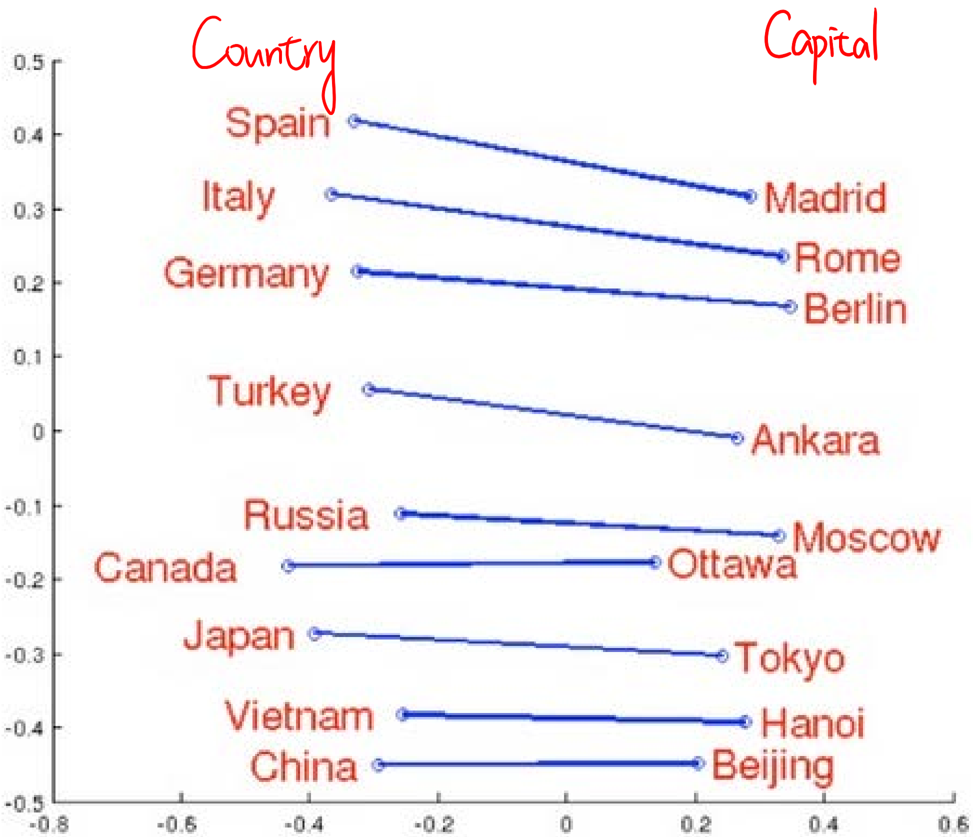
predicting the word given its context

- Skip-gram

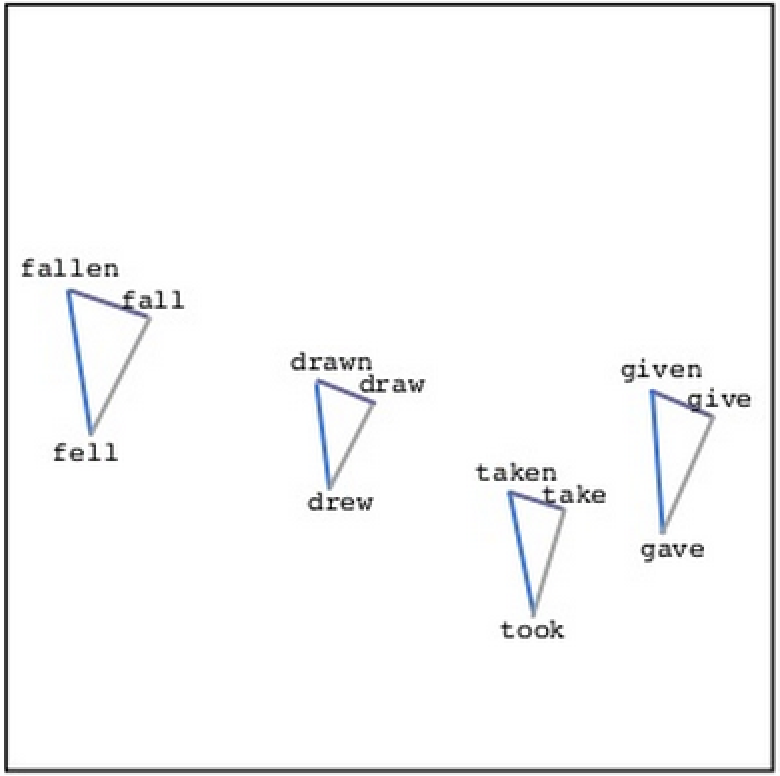


predicting the context given a word

Word Embedding

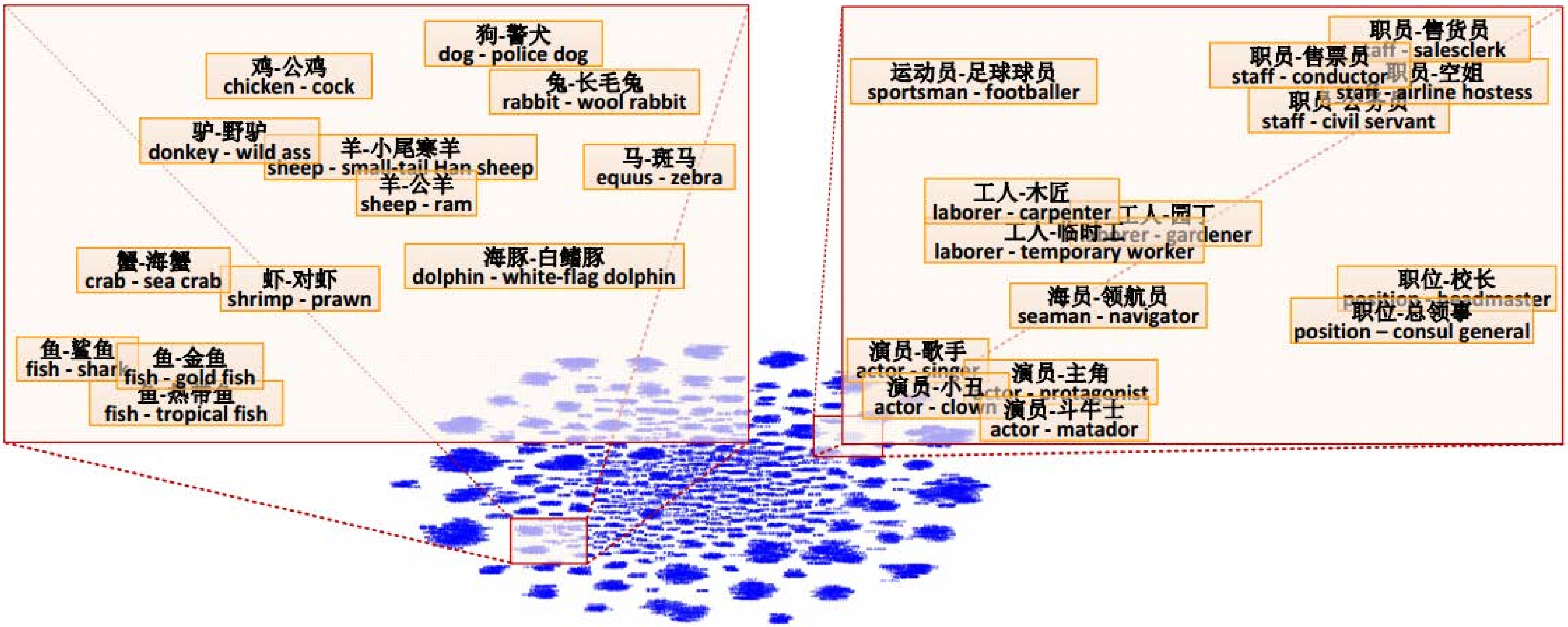


3 tenses of verb



Source: http://www.slideshare.net/hustwj/cikm-keynotenov2014

Word Embedding



Fu, Ruiji, et al. "Learning semantic hierarchies via word embeddings."*Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics: Long Papers*. Vol. 1. 2014.

Word Embedding

- Characteristics

$$V(Germany) \approx V(Berlin) - V(Rome) + V(Italy)$$

$$V(hotter) - V(hot) \approx V(bigger) - V(big)$$

$$V(Rome) - V(Italy) \approx V(Berlin) - V(Germany)$$

$$V(king) - V(queen) \approx V(uncle) - V(aunt)$$

- Solving analogies

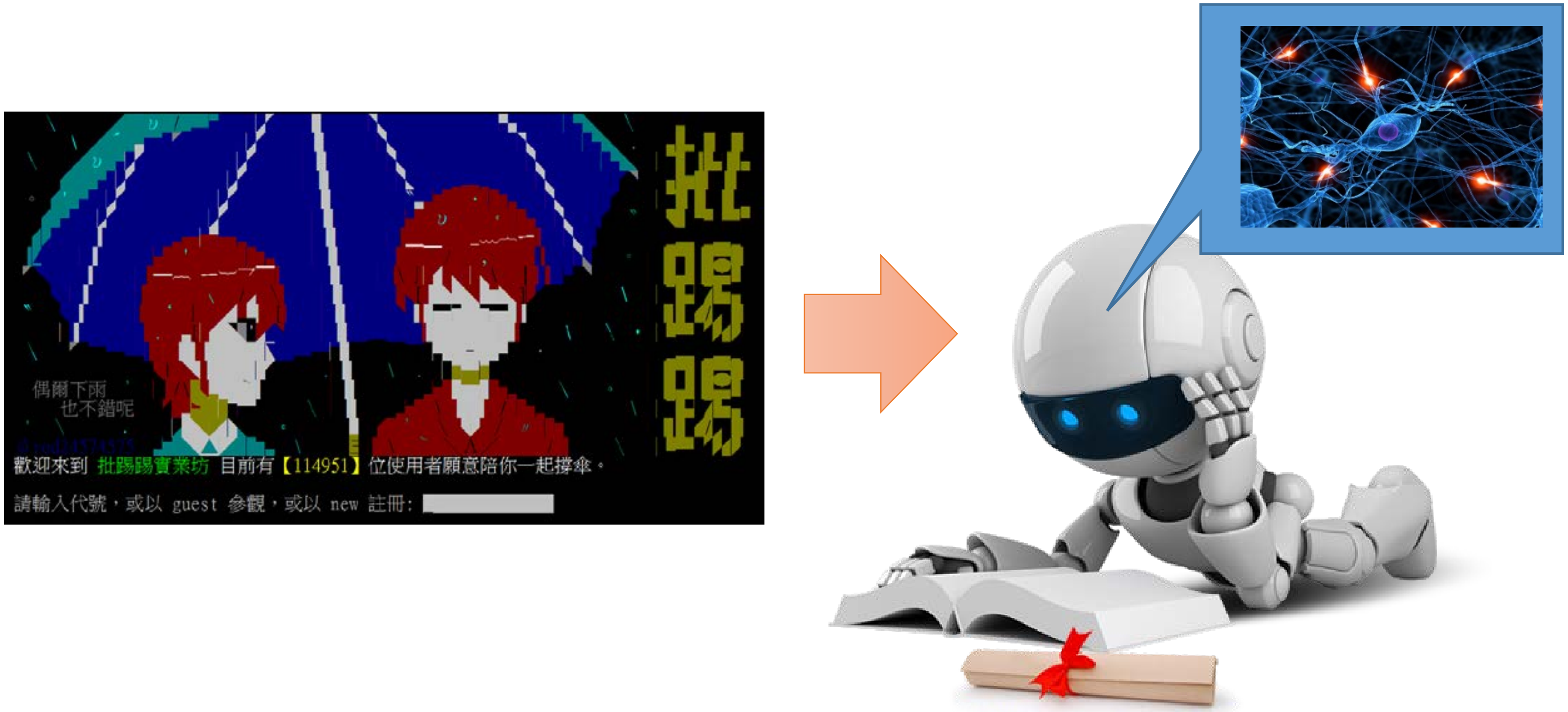
Rome : Italy = Berlin : ?

Compute $\underline{V(Berlin) - V(Rome) + V(Italy)}$

Find the word w with the closest V(w)

Demo

- Machine learns the meaning of words from reading a lot of documents without supervision

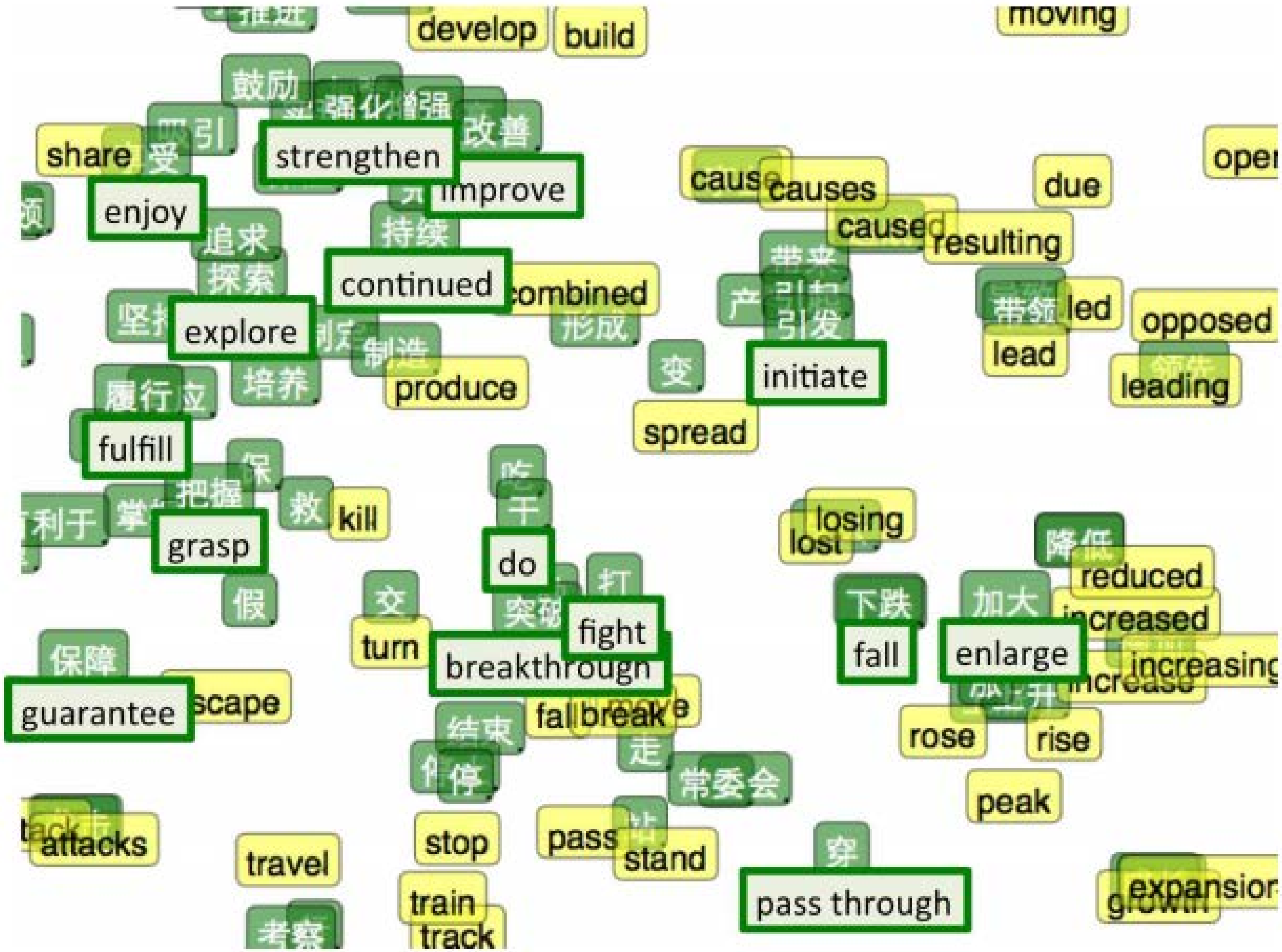


Demo

- Model used in demo is provided by 陳仰德
  - Part of the project done by 陳仰德、林資偉
  - TA: 劉元銘
  - Training data is from PTT (collected by 葉青峰)

English word: Green ones mean we already know its Chinese meaning.
Yellow ones mean we don't know its Chinese meaning.

Multi-lingual Embedding



Bilingual Word Embeddings for Phrase-Based Machine Translation, Will Zou, Richard Socher, Daniel Cer and Christopher Manning, EMNLP, 2013

Document Embedding

Various length.

Use count.

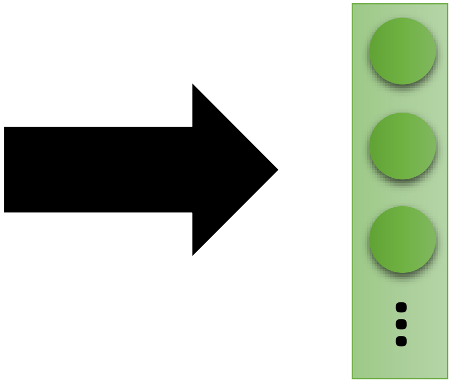
- word sequences with different lengths → the vector with the same length

embedding

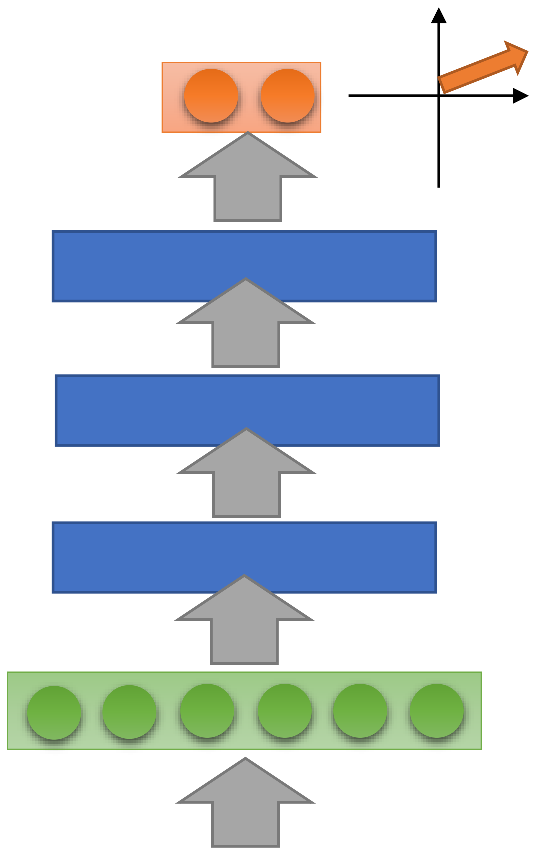
  - The vector representing the meaning of the word sequence
  - A word sequence can be a document or a paragraph

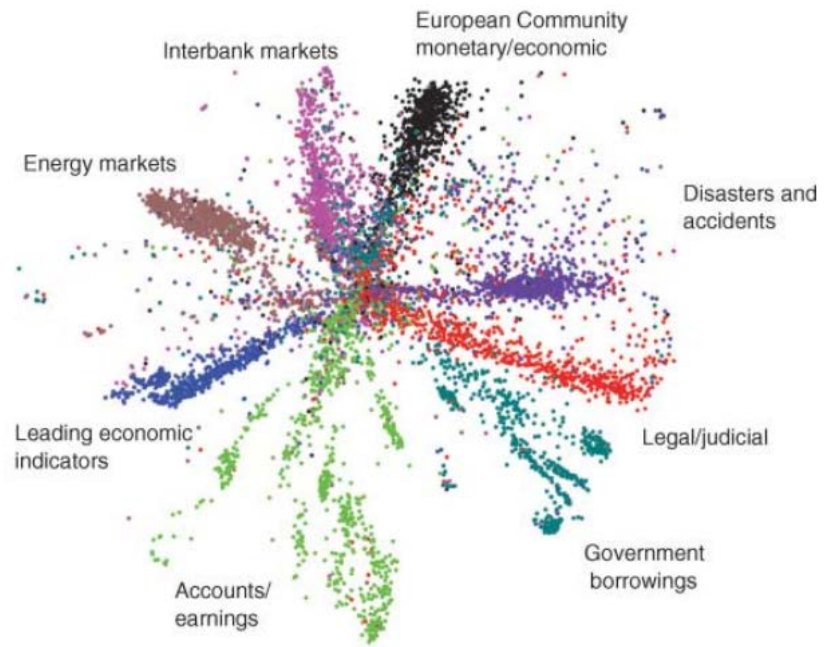
word sequence

(a document or paragraph)



Semantic Embedding





Bag-of-word

Reference: Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507

Just record the count of the words. (Ignore the order.)

Beyond Bag of Word

- To understand the meaning of a word sequence, the order of the words can not be ignored.

References on the next page.

white blood cells destroying an infection → positive

exactly the same bag-of-word

different meaning

an infection destroying white blood cells → negative

Beyond Bag of Word

- ***Paragraph Vector***: Le, Quoc, and Tomas Mikolov. "Distributed Representations of Sentences and Documents.“ ICML, 2014
- ***Seq2seq Auto-encoder***: Li, Jiwei, Minh-Thang Luong, and Dan Jurafsky. "A hierarchical neural autoencoder for paragraphs and documents." arXiv preprint, 2015
- ***Skip Thought***: Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, Sanja Fidler, “Skip-Thought Vectors” arXiv preprint, 2015.