Unsupervised Learning: Word Embedding

Word Embedding

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

1-of-N Encoding

apple = [1 0 0 0 0]

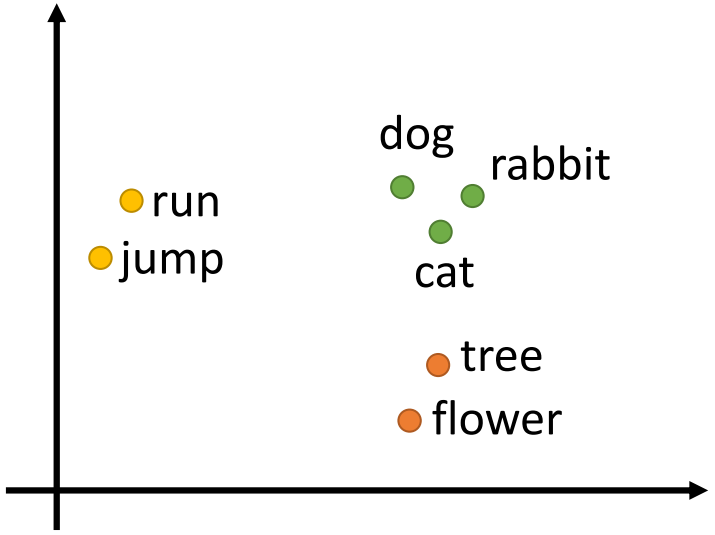
bag = [0 1 0 0 0]

cat = [0 0 1 0 0]

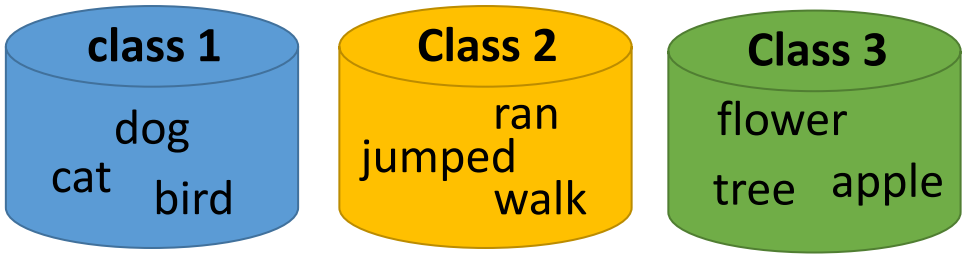
dog = [0 0 0 1 0]

elephant = [0 0 0 0 1]

Word Embedding



Word Class



蔡英文、馬英九 are something very similar

You shall know a word by the company it keeps

馬英九 520宣誓就職

蔡英文 520宣誓就職

Word Embedding

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

How to exploit the context?

- **Count based**
  - If two words w_i and w_j frequently co-occur, $V(w_i)$ and $V(w_j)$ would be close to each other

$V(w_i) \cdot V(w_j)$

Inner product

$N_{i,j}$

Number of times w_i and w_j in the same document

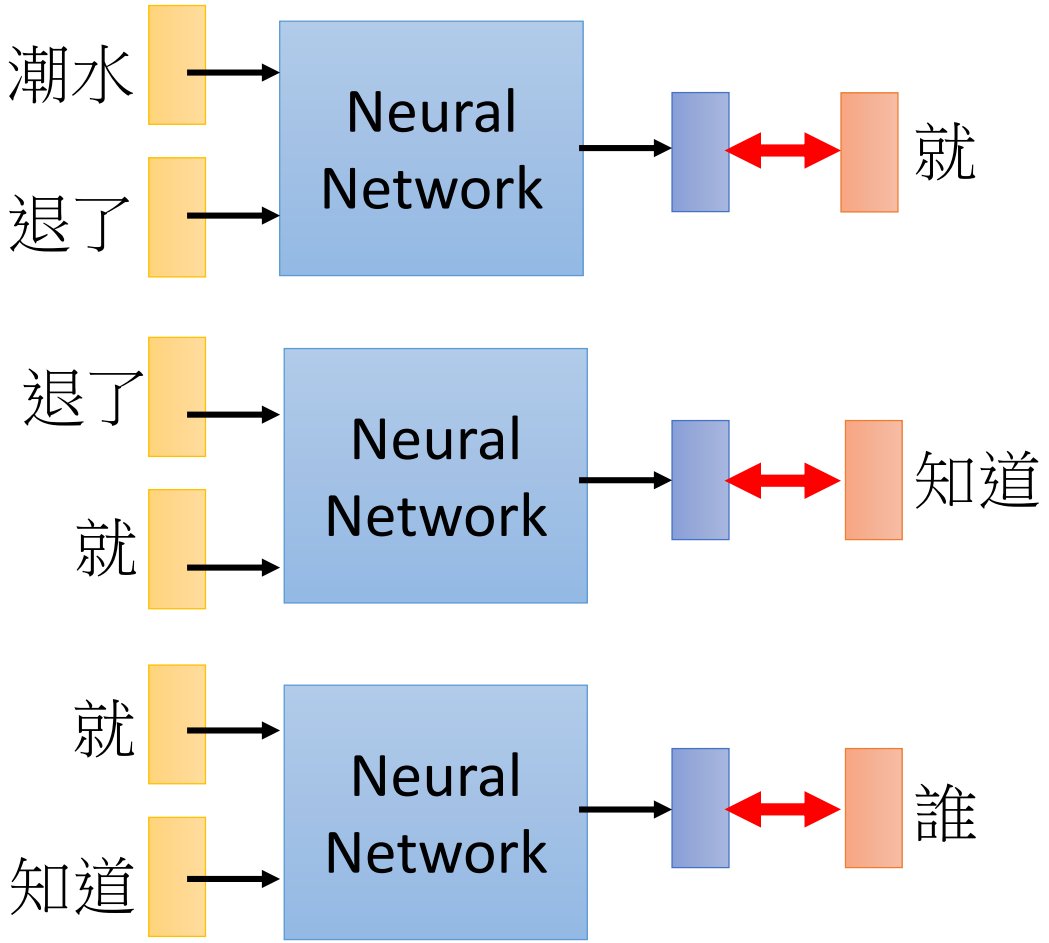
- **Perdition based**

Prediction-based – Training

Collect data:

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

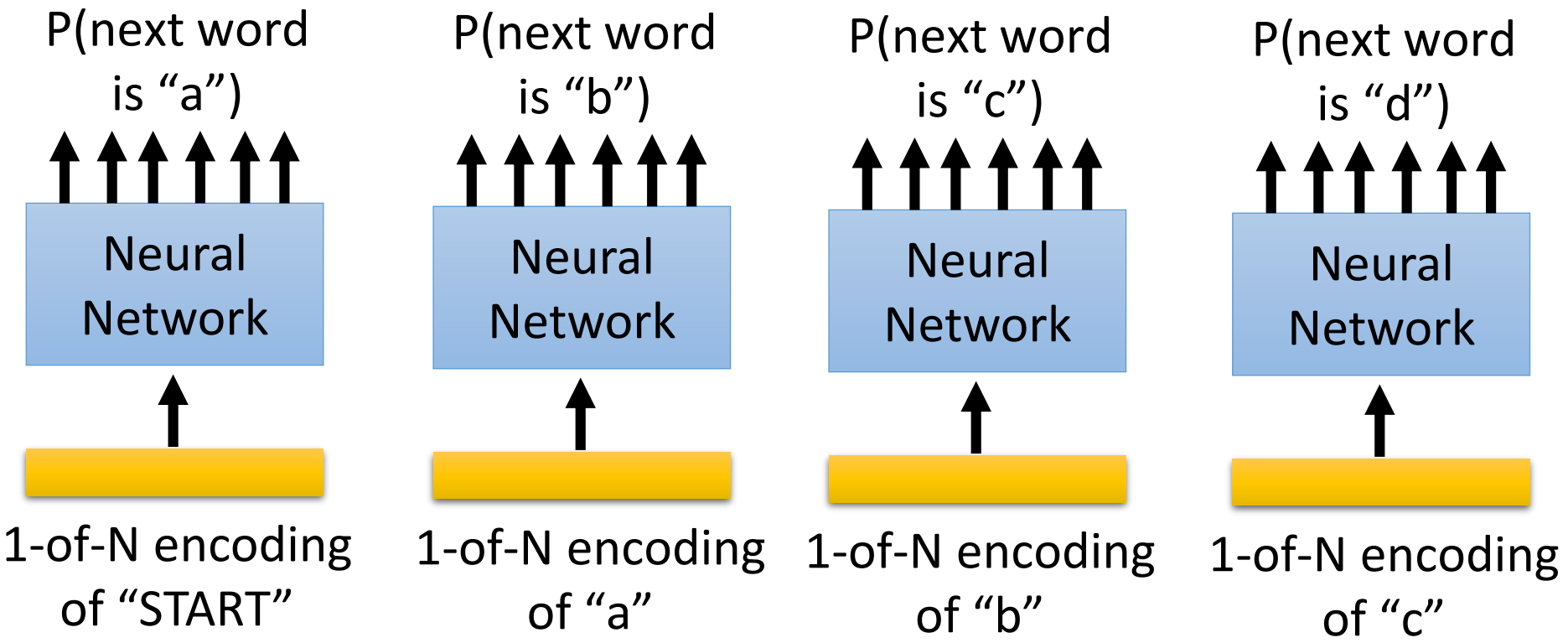
Minimizing cross entropy



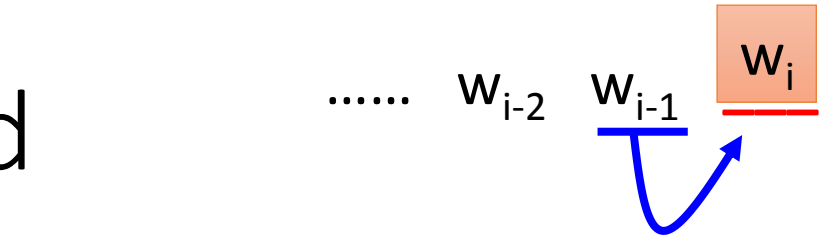
Prediction-based – Language Modeling

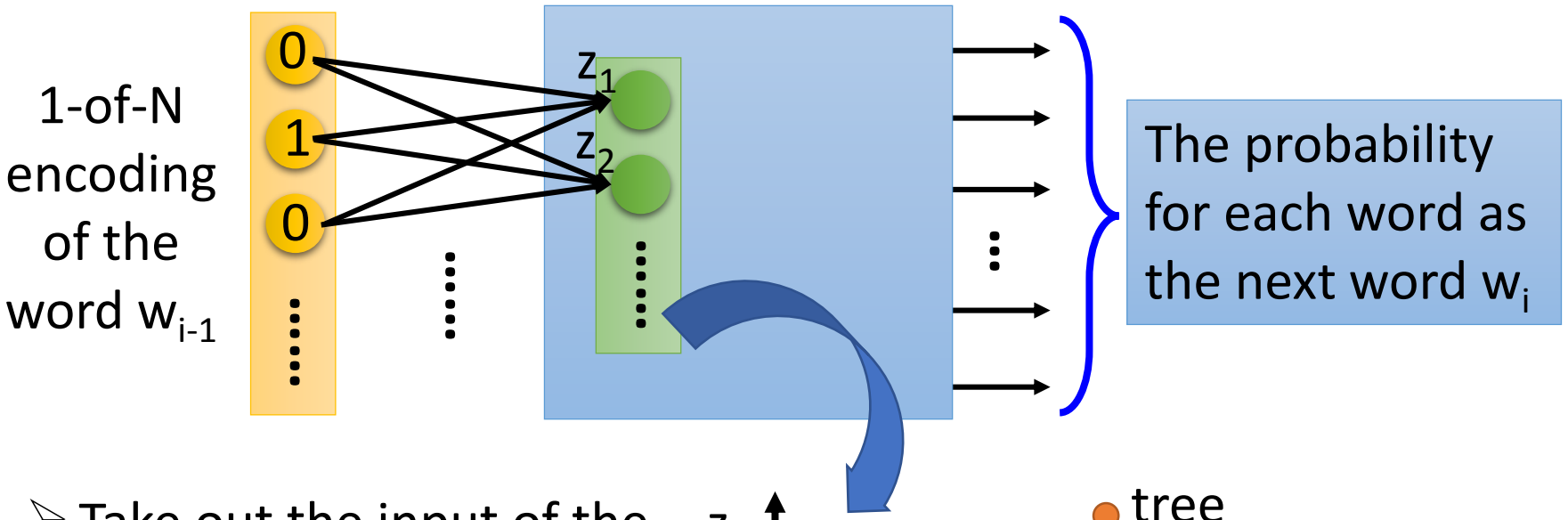
P("a b c d")=P(a|START)P(b|a)P(c|b)P(d|c)

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

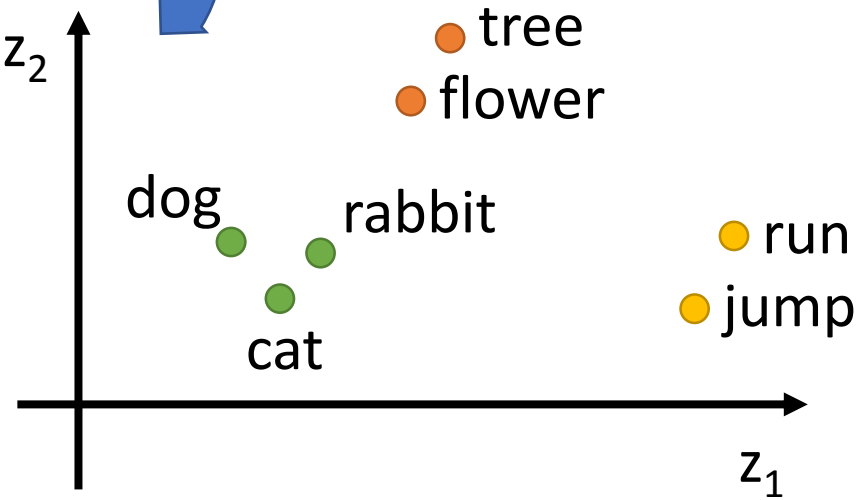


Prediction-based

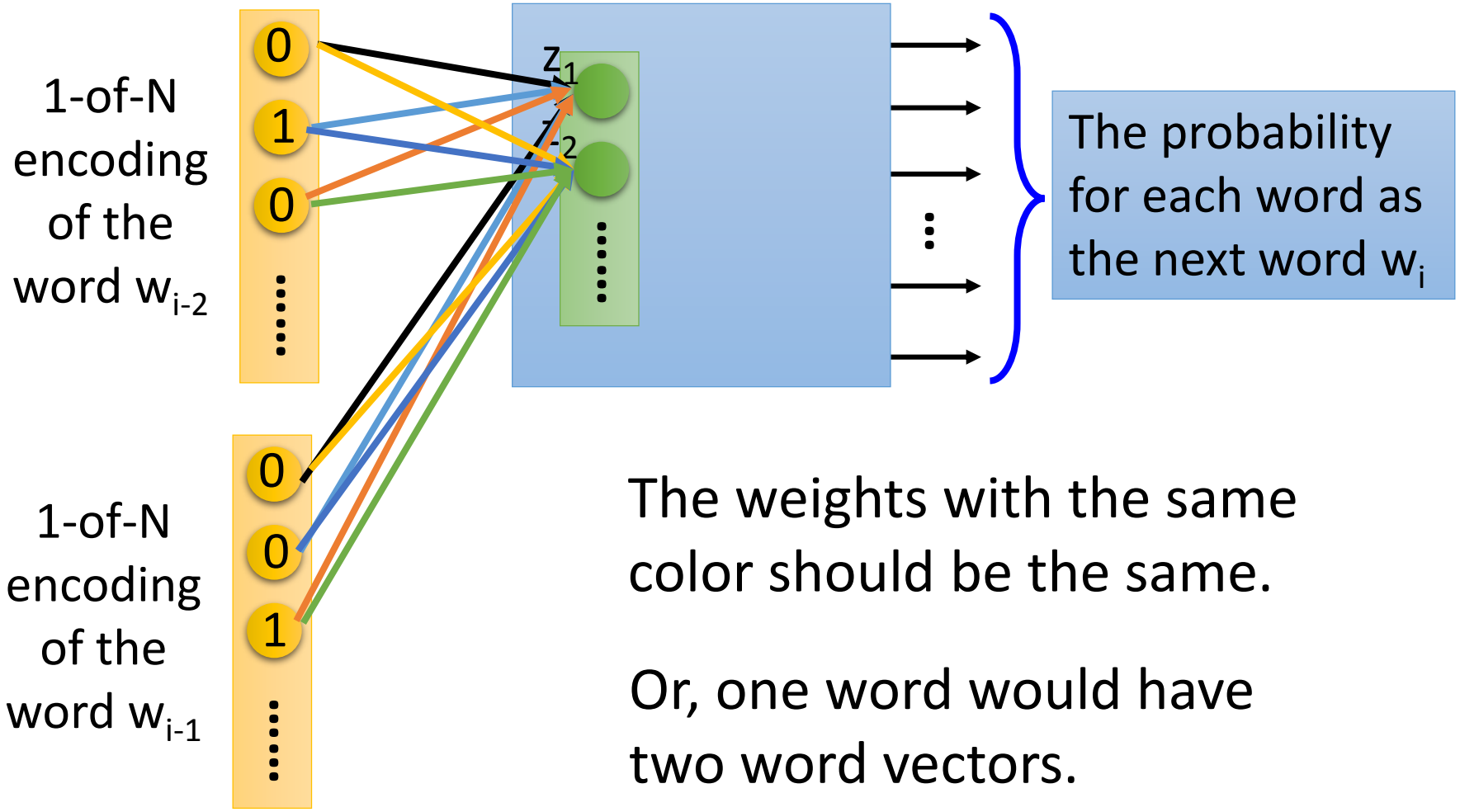




- Take out the input of the neurons in the first layer
- Use it to represent a word w
- Word vector, word embedding feature: V(w)



Prediction-based–Sharing Parameters

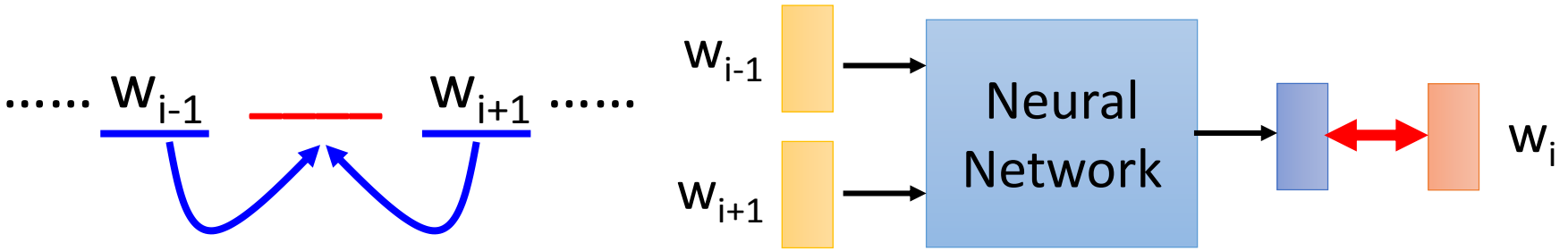


The weights with the same color should be the same.

Or, one word would have two word vectors.

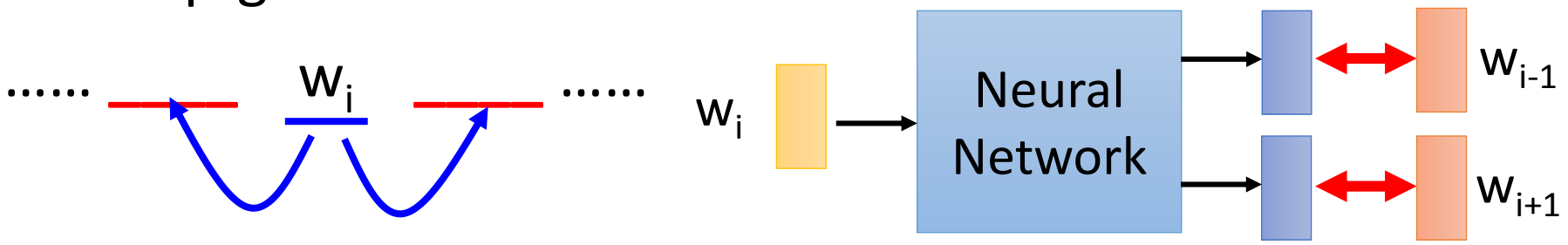
Prediction-based–Various Architectures

- Continuous bag of word (CBOW) model



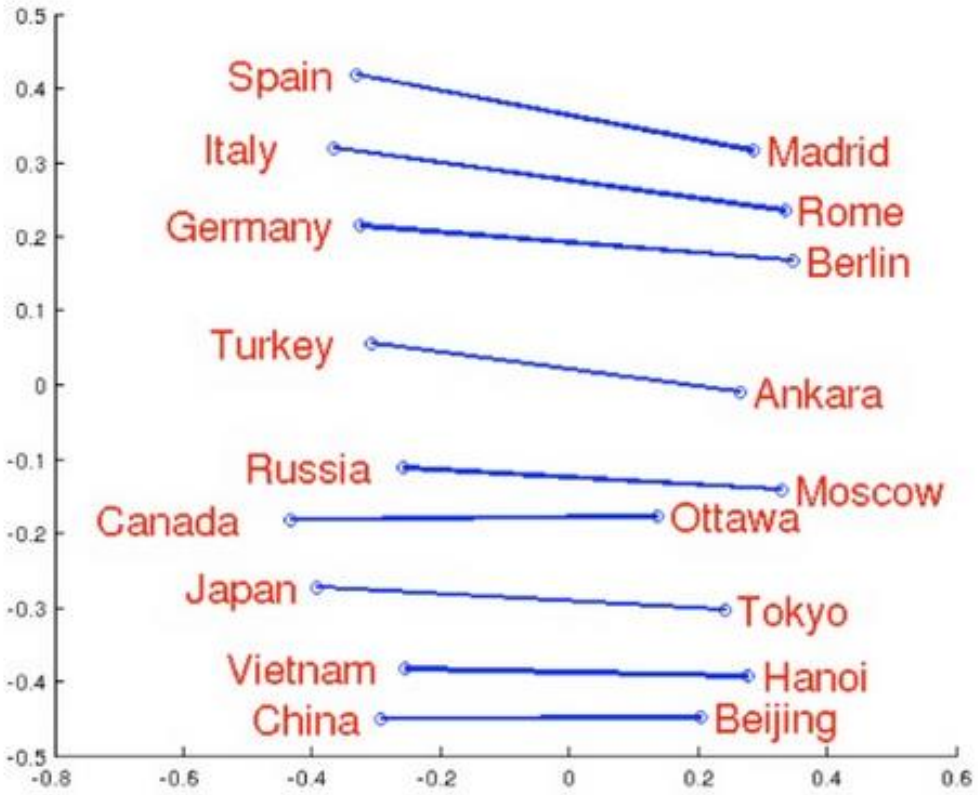
predicting the word given its context

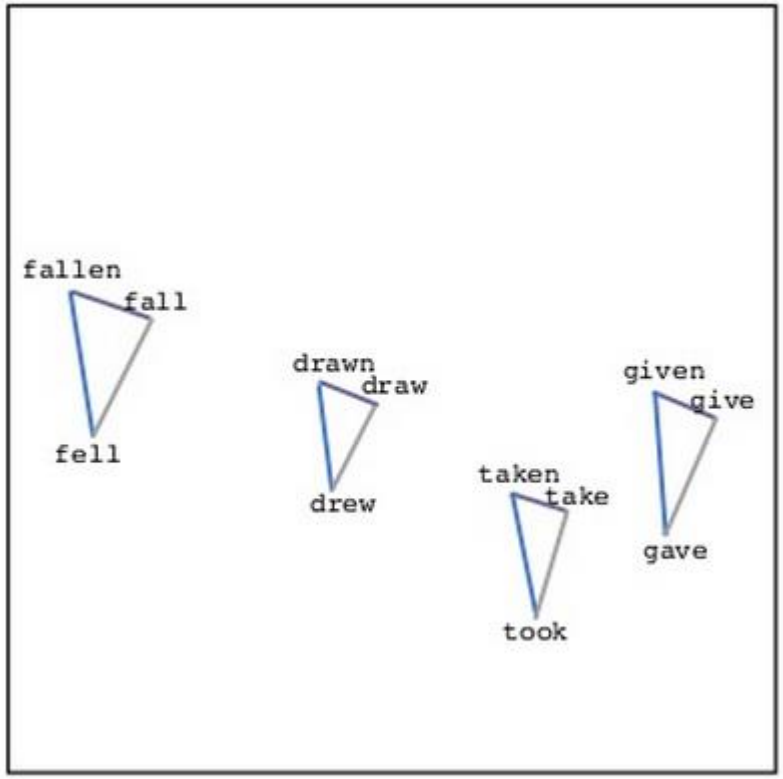
- Skip-gram



predicting the context given a word

Word Embedding





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

Word Embedding

- Characteristics

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

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

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

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

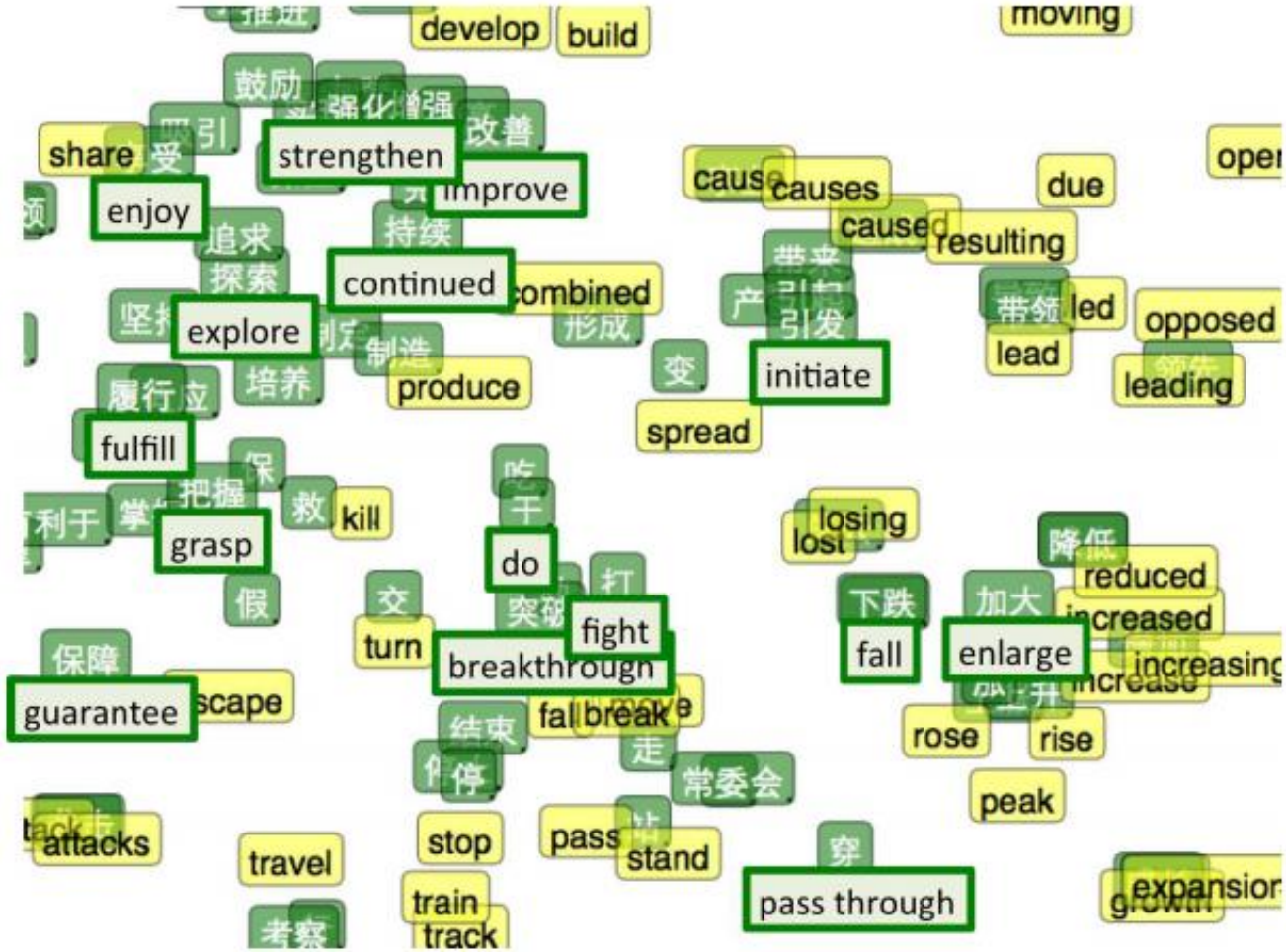
- Solving analogies

Rome : Italy = Berlin : ?

Compute $V(Berlin) - V(Rome) + V(Italy)$

Find the word w with the closest V(w)

Multi-lingual Embedding



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