Deep Learning
Do machines know the meaning of a word?

Hung-yi Lee

Language Technology

spam detection

(http://spam-filter-review.toptenreviews.com/)

Part-of-speech Tagging

John saw the saw.

John	saw	the	saw.
PN	V	D	N

Name Entity Recognition

這 位 是 李 宏 毅

Name of People

Sentiment Analysis

這部電影太糟了

Negative (負雷)

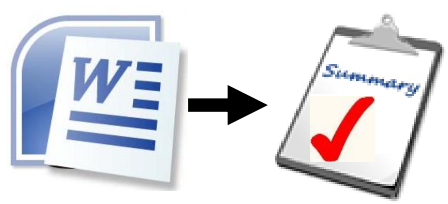
Translation

"Machine learning ……"

⬍

"機器學習 ……"

Summarization



document summary

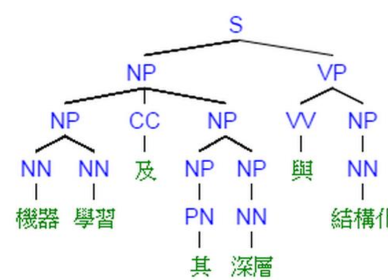
Retrieval

Google 台灣

Speech Recognition

大家好……

Syntactic Analysis

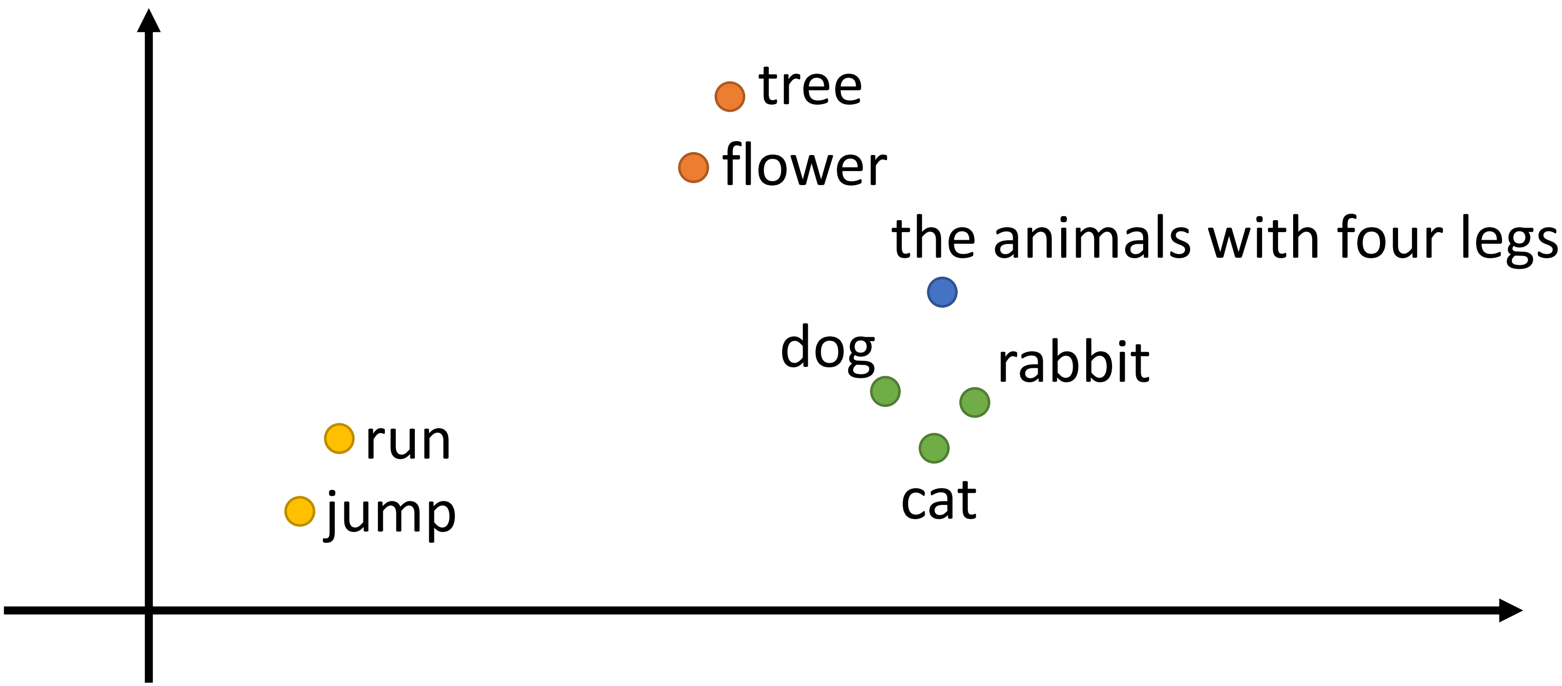


Do machine really understand human language?

http://cse3521.artifice.cc/chinese-room.html

Meaning Representation

Do machine know the meaning of a word or word sequence?



Meaning of Word

Predicting the next word

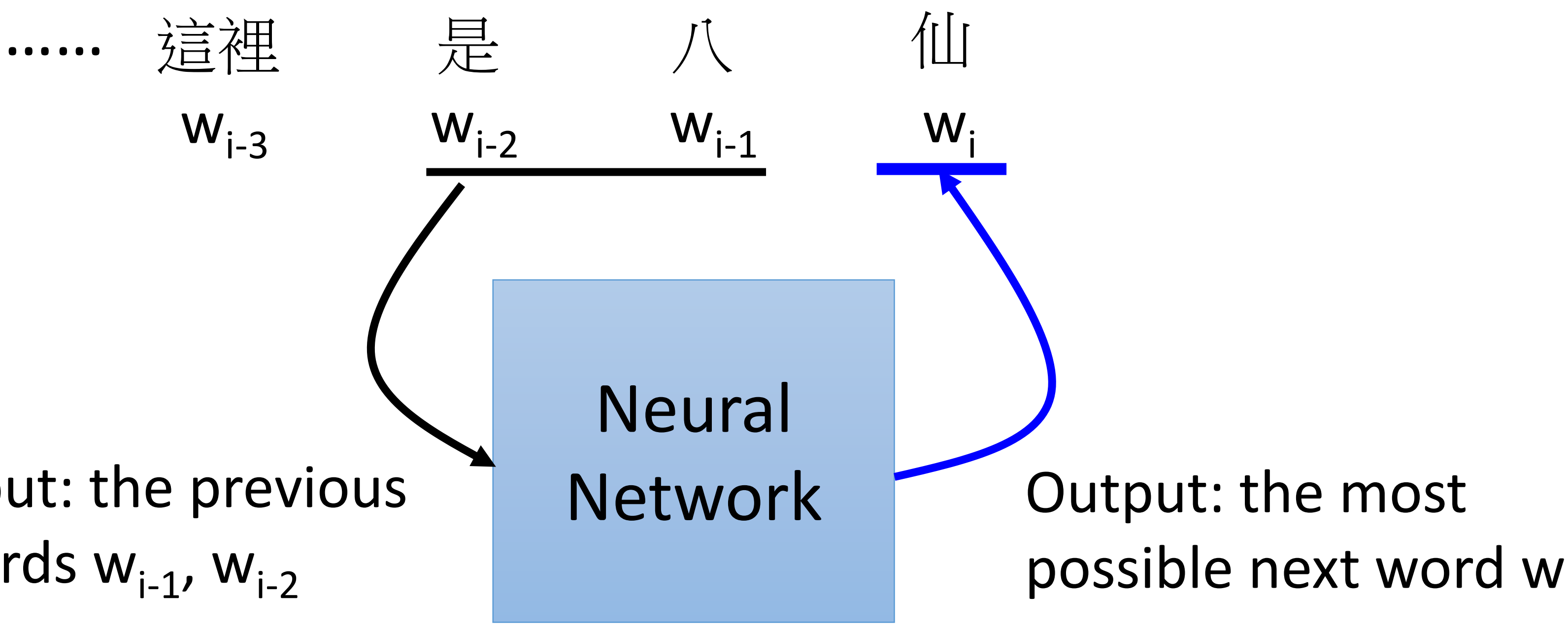
- Given a sequence of words, predict the next word

麻煩這系列的請到政黑或其他地方討論好嗎?這裡是八 04/27 00:40

Ref: http://pttpedia.pixnet.net/blog/post/167961207-%E9%80%99%E8%A3%A1%E6%98%AF%E5%85%AB%E4%BB%99%E6%A8%82%E5%9C%92

Predicting the next word

- Given a sequence of words, predict the next word



Each word should be represented as a feature vector.

Predicting the next word

1-of-N Encoding

lexicon = {apple, bag, cat, dog, elephant}

apple = [1 0 0 0 0]

bag = [0 1 0 0 0]

cat = [0 0 1 0 0]

dog = [0 0 0 1 0]

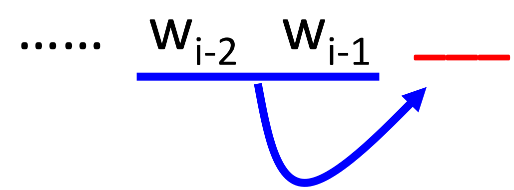
elephant = [0 0 0 0 1]

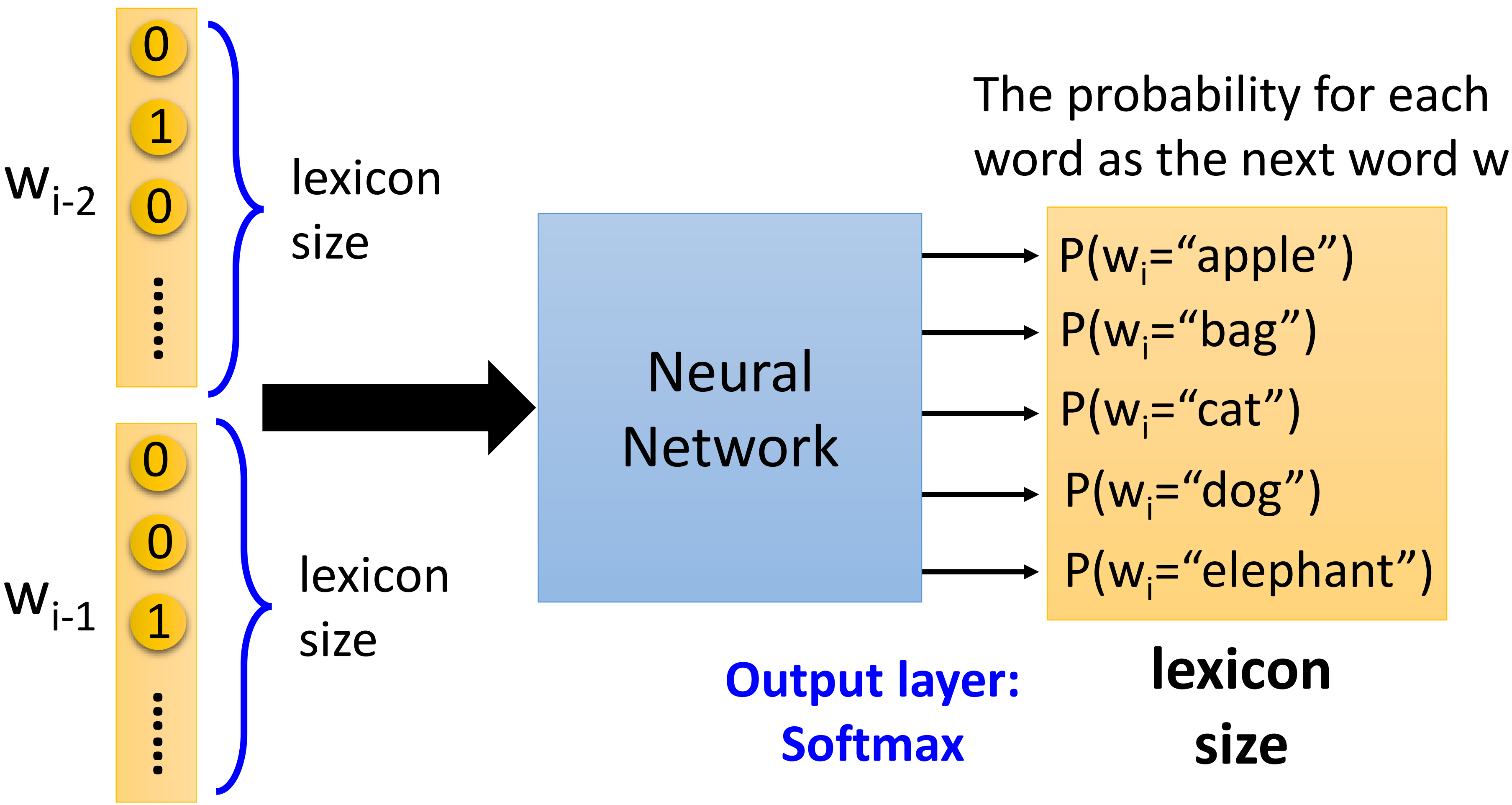
The vector is lexicon size.

Each dimension corresponds to a word in the lexicon

The dimension for the word is 1, and others are 0

Predicting the next word





Predicting the next word

Application?

- Training:

Collect data:

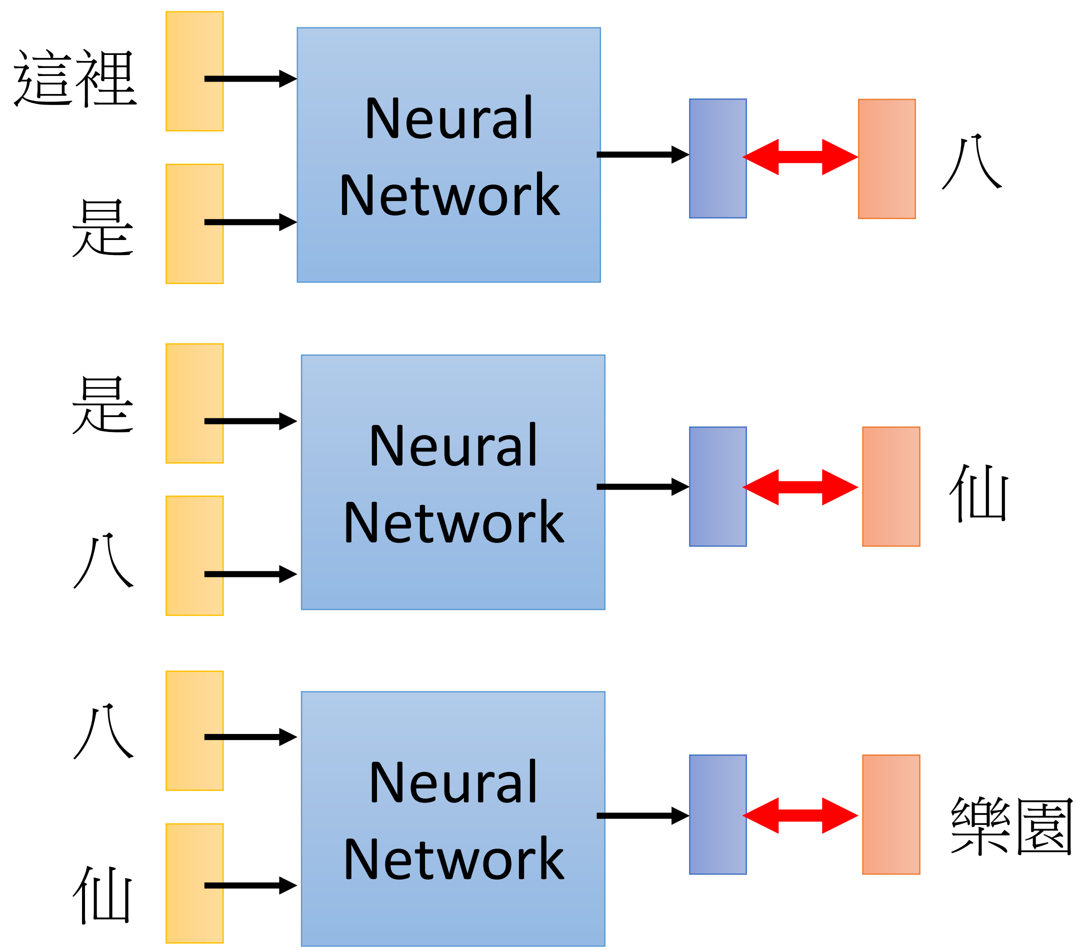
這裡 是 八 仙 樂園

………

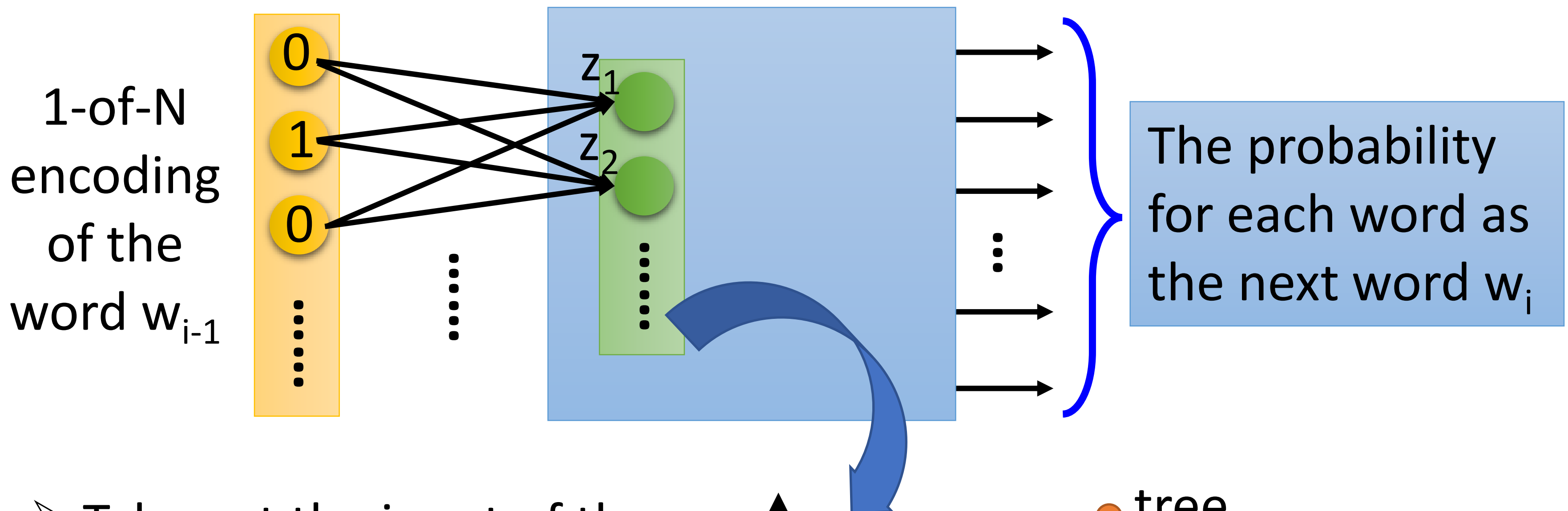
………

………

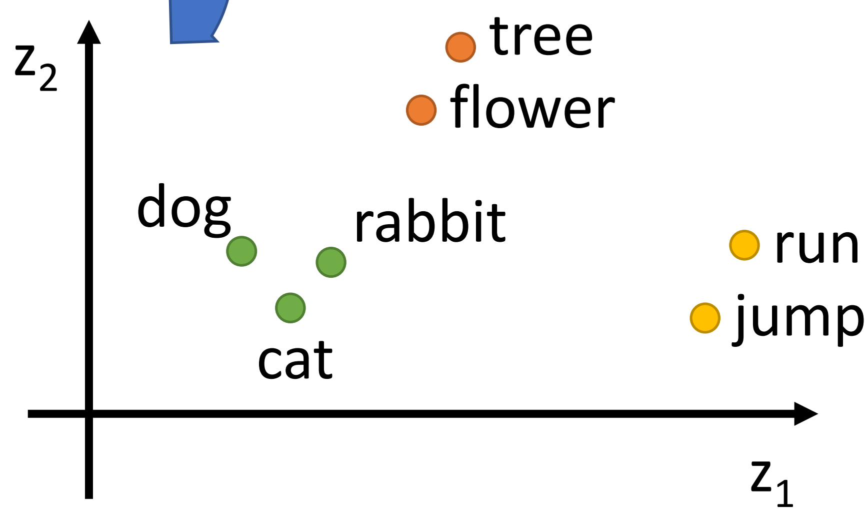
Minimizing cross entropy



Word Vector

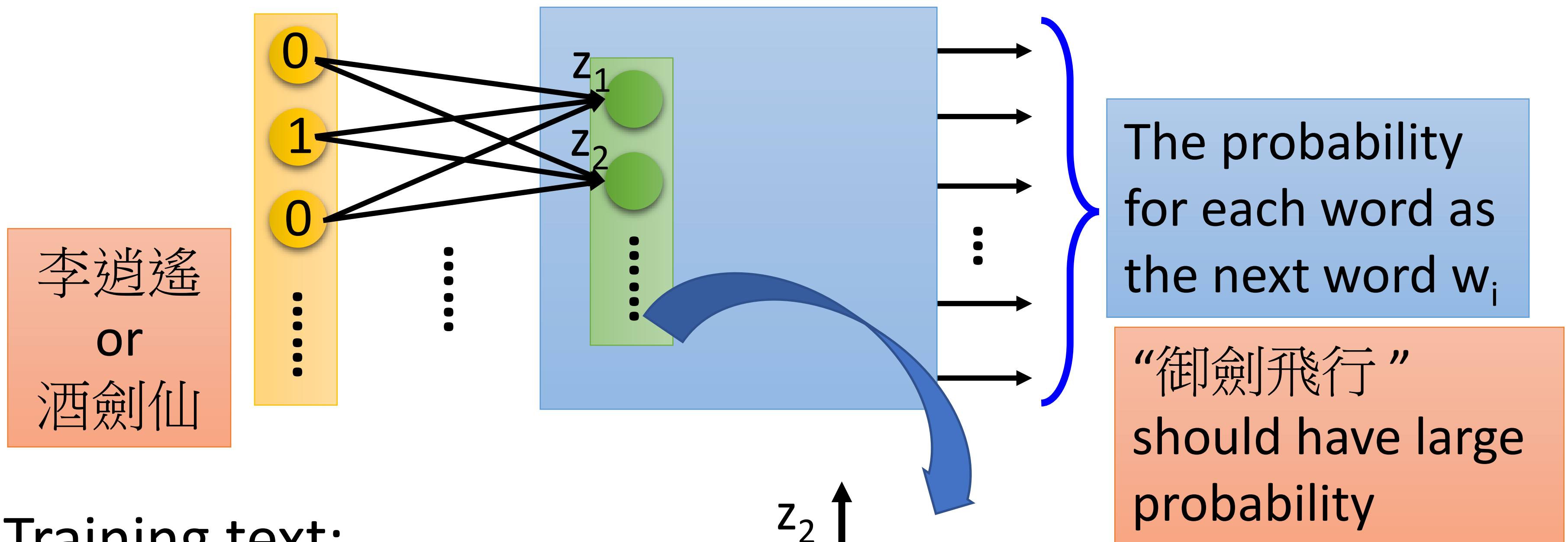


- Take out the input of the neurons in the first layer
- Use it to represent a word w
- Word vector, word embedding feature: V(w)



Word Vector

You shall know a word by the company it keeps



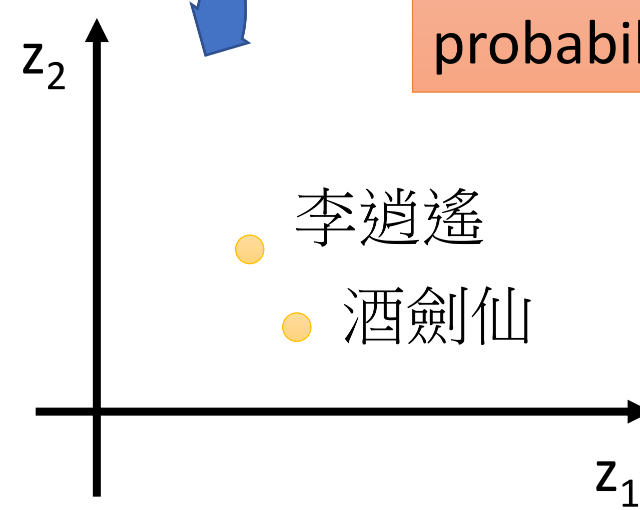
Training text:

…… 李逍遙 御劍飛行 ……

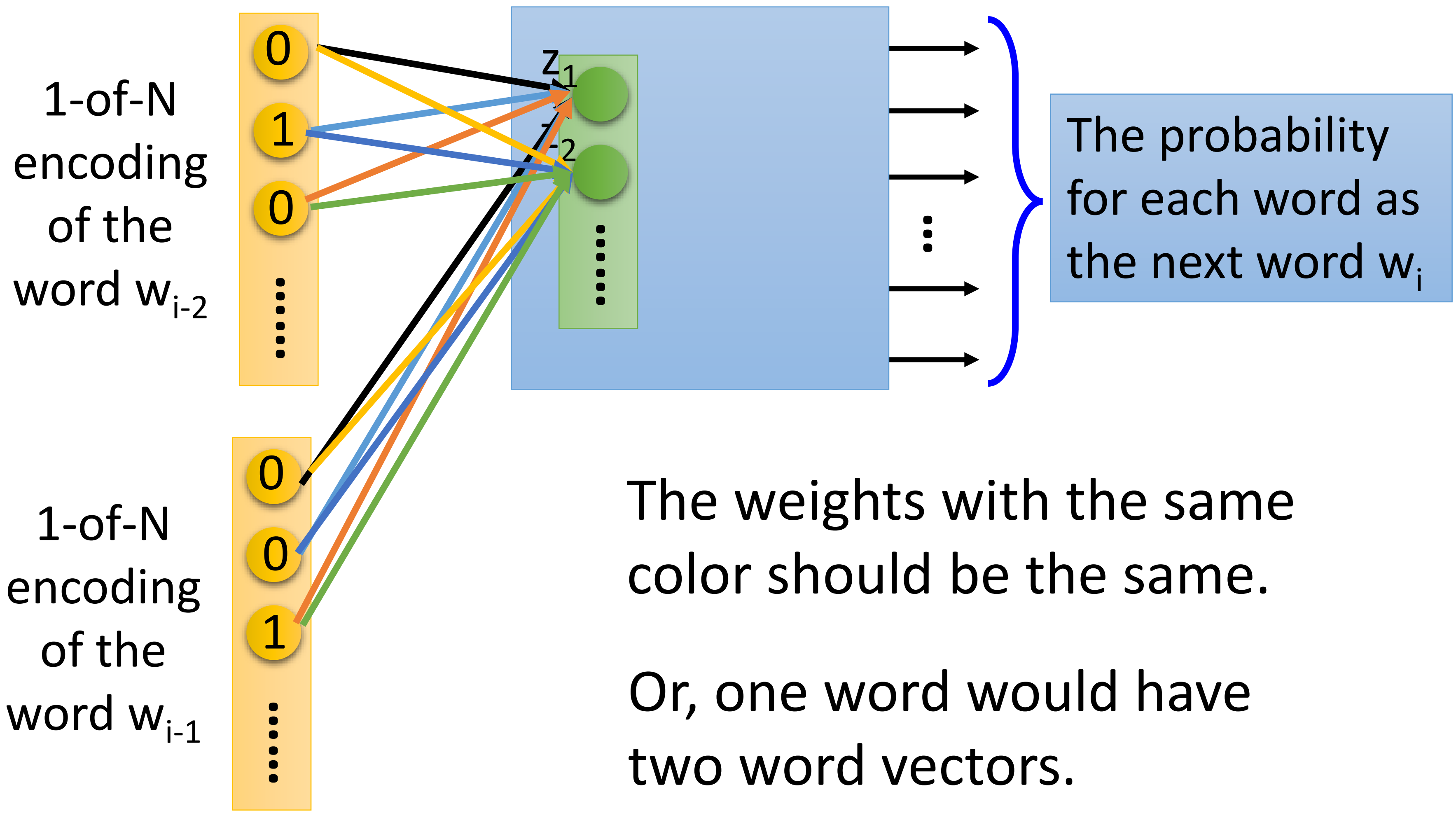
w_{i-1} w_i

…… 酒劍仙 御劍飛行 ……

w_{i-1} w_i



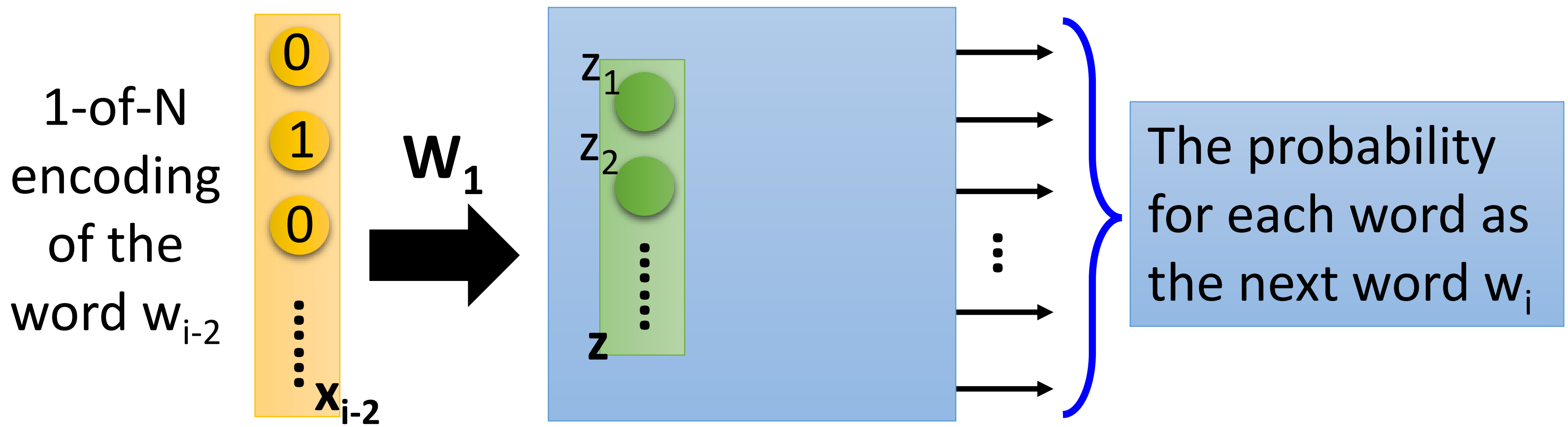
Word Vector – Sharing Parameters

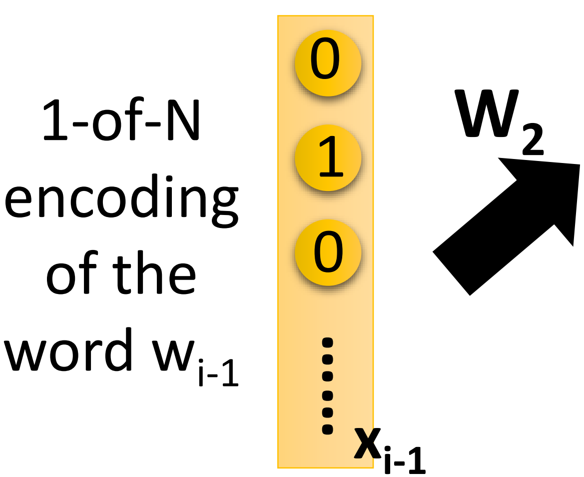


The weights with the same color should be the same.

Or, one word would have two word vectors.

Word Vector – Sharing Parameters





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

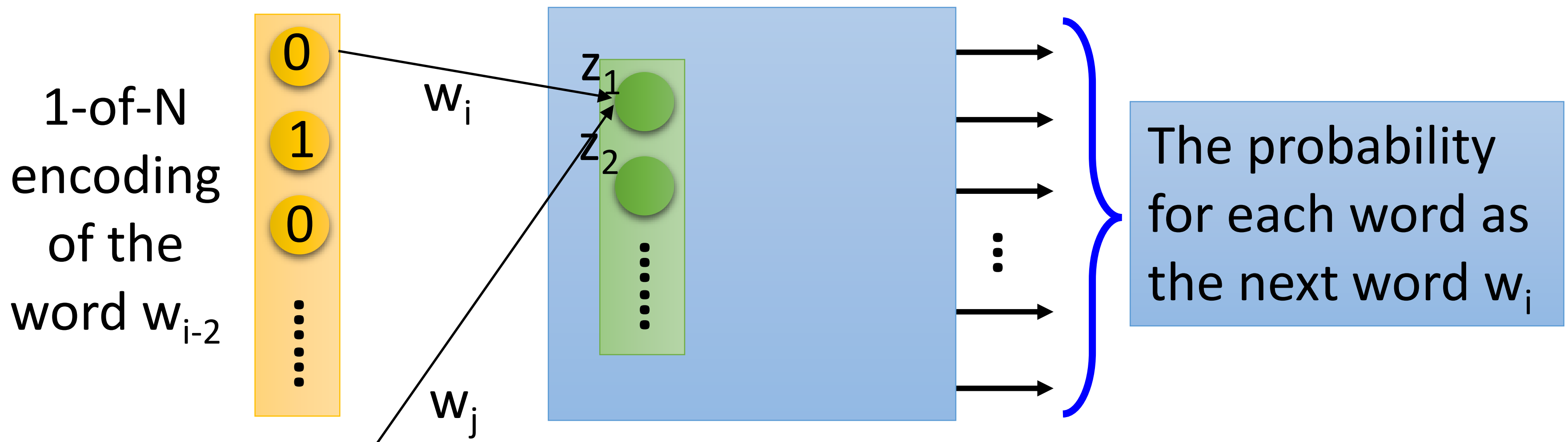
The length of $\mathbf{z}$ is $|Z|$.

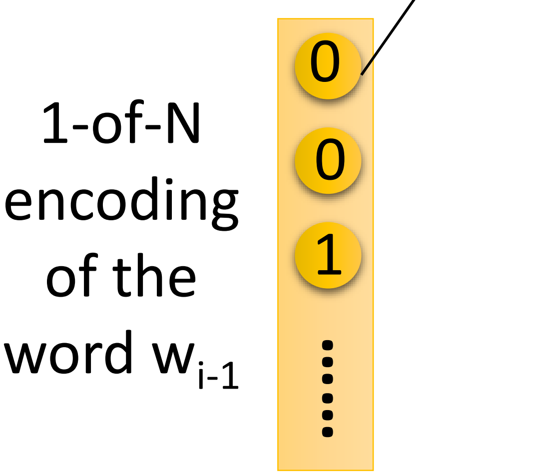
$$\mathbf{z} = \mathbf{W}_1\,\mathbf{x}_{i-2} + \mathbf{W}_2\,\mathbf{x}_{i-1}$$

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

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

Word Vector – Sharing Parameters





How to make w_i equal to w_j

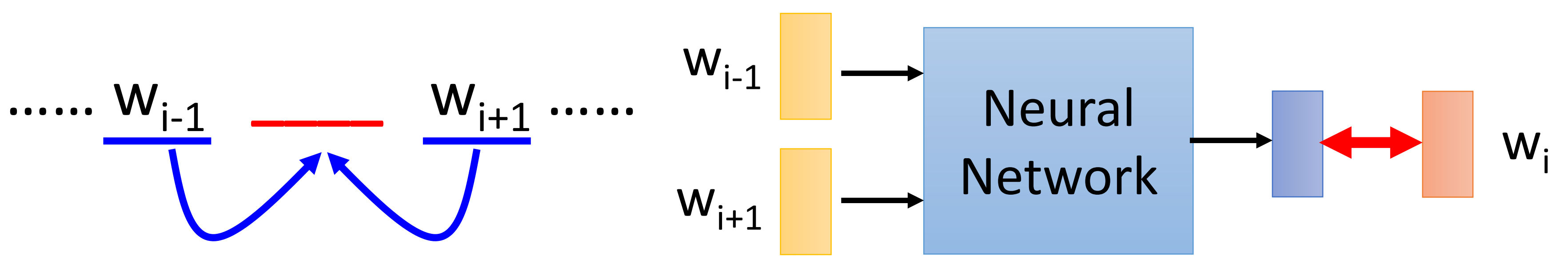
Given w_i and w_j the same initialization

$$w_i \leftarrow w_i - \eta \frac{\partial C}{\partial w_i} - \eta \frac{\partial C}{\partial w_j}$$

$$w_j \leftarrow w_j - \eta \frac{\partial C}{\partial w_j} - \eta \frac{\partial C}{\partial w_i}$$

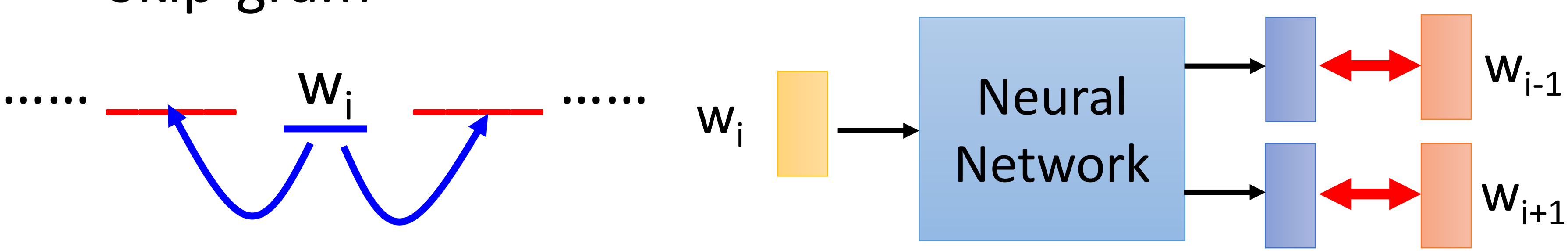
Word Vector
– Various Architectures

- Continuous bad of word (CBOW) model



predicting the word given its context

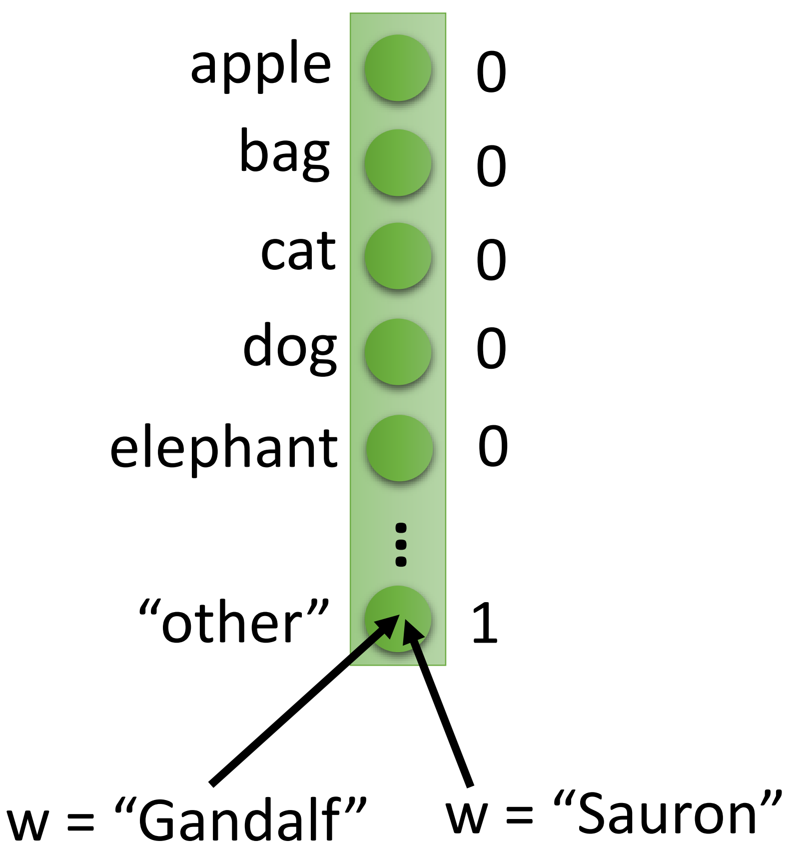
- Skip-gram



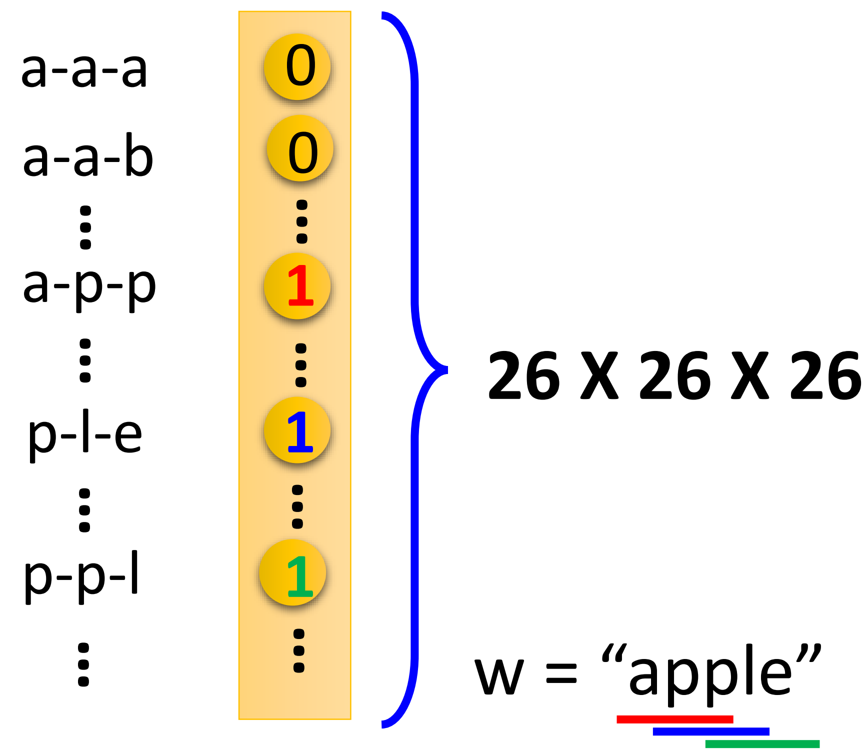
predicting the context given a word

Beyond 1-of-N encoding

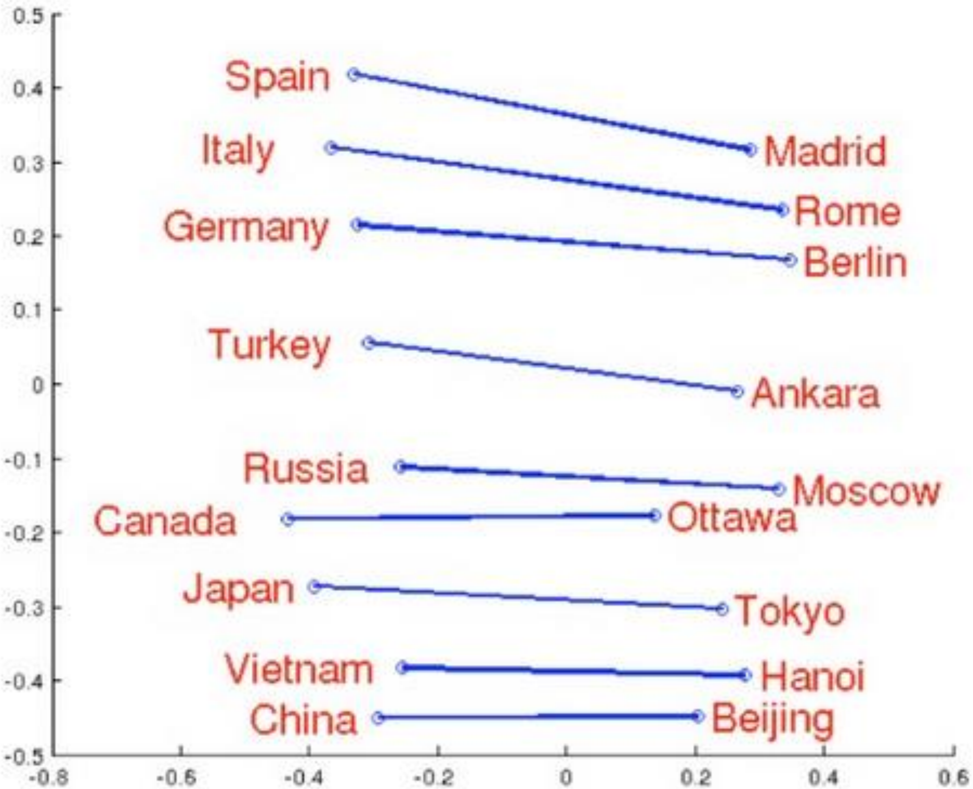
Dimension for "Other"

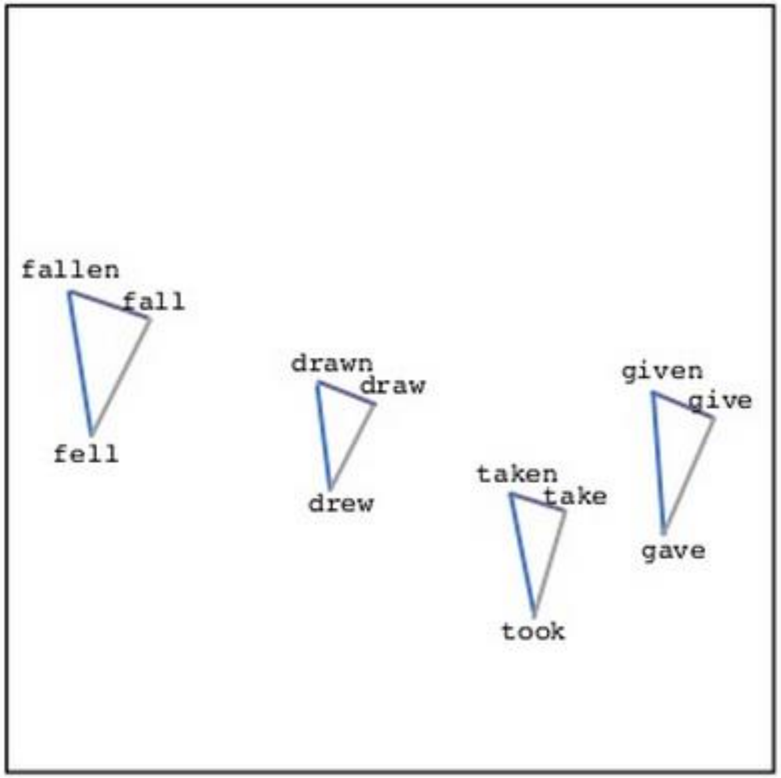


Word hashing



Word Vector





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

Word Vector

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

- Characteristics

$$V(hotter) - V(hot) \approx V(bigger) - V(big)$$
$$V(Rome) - V(Italy) \approx V(Berlin) - V(Germany)$$
$$V(king) - V(queen) \approx V(uncle) - V(aunt)$$

- Solving analogies

Rome : Italy = Berlin : ?

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

Find the word w with the closest V(w)

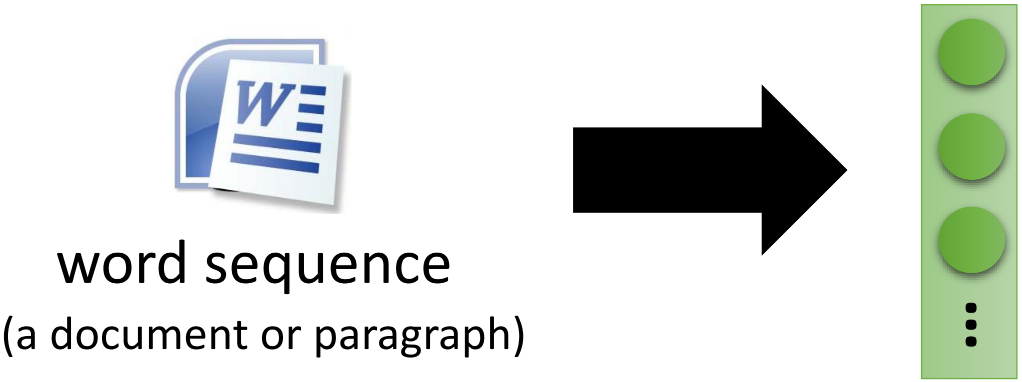
Demo

- 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)

Meaning of Word Sequence

Meaning of Word Sequence

- word sequences with different lengths → the vector with the same length
  - The vector representing the meaning of the word sequence
  - A word sequence can be a document or a paragraph

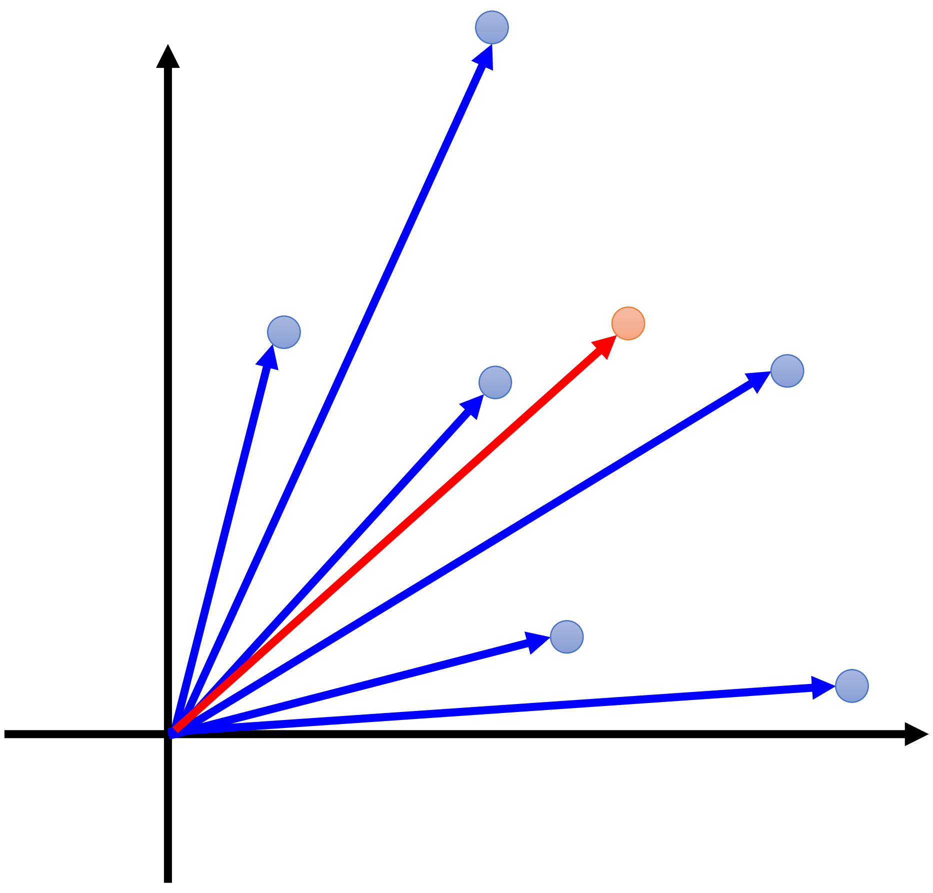


Meaning of Word Sequence - Outline

Deep Structured Semantic Model (DSSM)	• Application: Information Retrieval (IR)
Recursive Deep Model	• Application: Sentiment Analysis
Paragraph Vector	• Unsupervised

Reference: http://www.msr-waypoint.net/pubs/198202/cikm2013_DSSM_fullversion.pdf

Information Retrieval (IR)



<u>Vector Space Model</u>

The documents are vectors in the space.

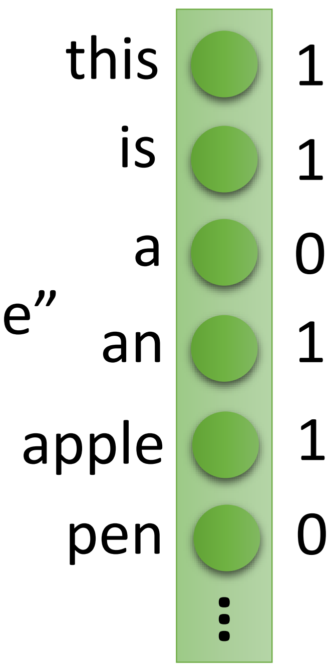
The query is also a vector.

How to use a vector to represent word sequences

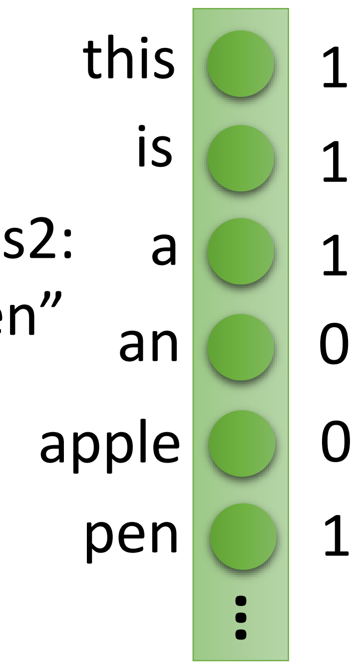
Information Retrieval (IR)

Bag-of-word

word string s1:
"This is an apple"



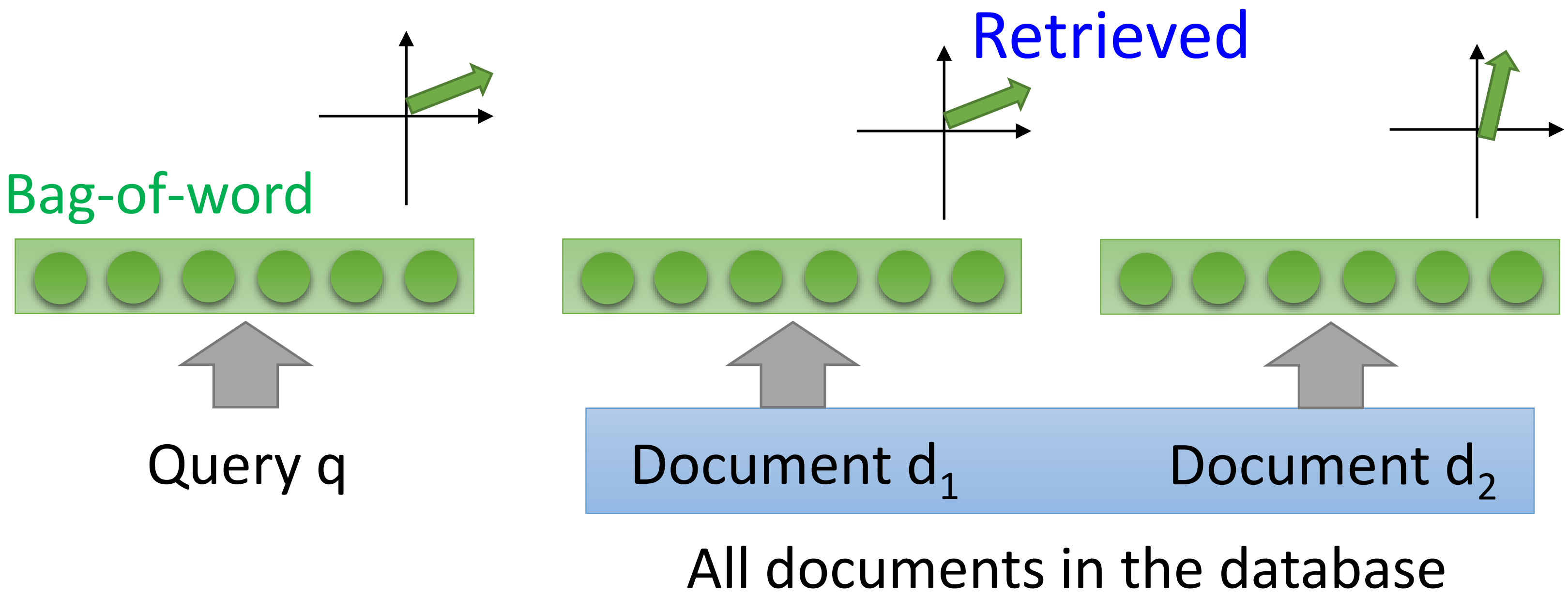
word string s2:
"This is a pen"



Weighted by IDF

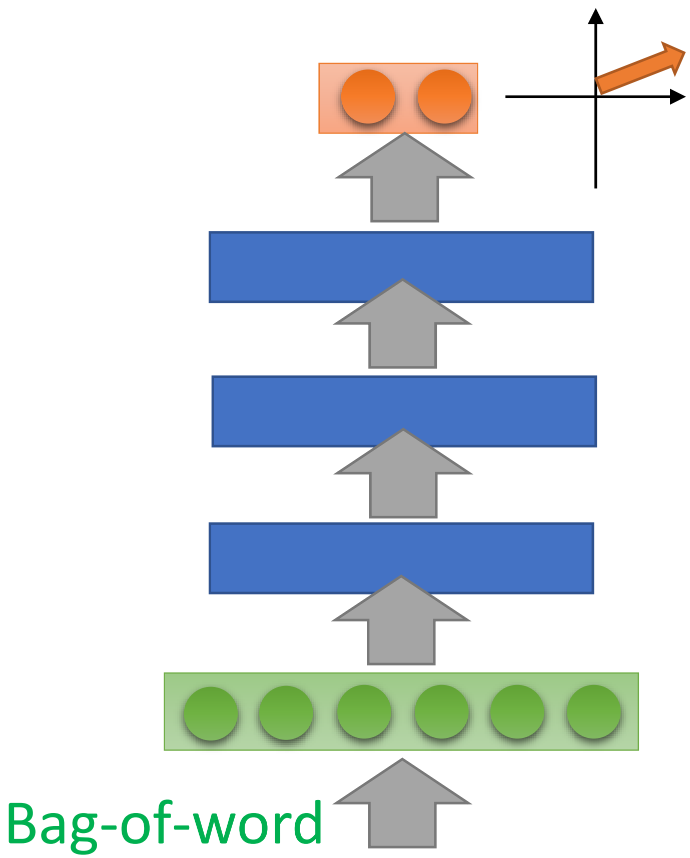
Information Retrieval (IR)

Vector Space Model + Bag-of-word

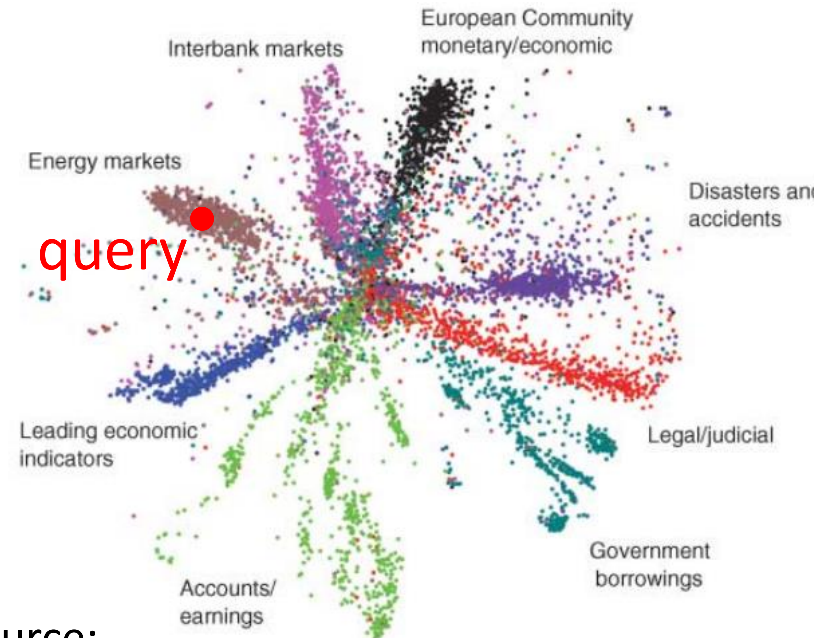


- All the words are treated as discrete tokens
- Never consider different words can have the same meaning, and the same word can have different meanings

IR - Semantic Embedding



word string
(document or query)



Source:
http://www.cs.toronto.edu/~hinton/science.pdf

How to achieve that? (No target ……)

Auto-encoder is one solution,
but not today

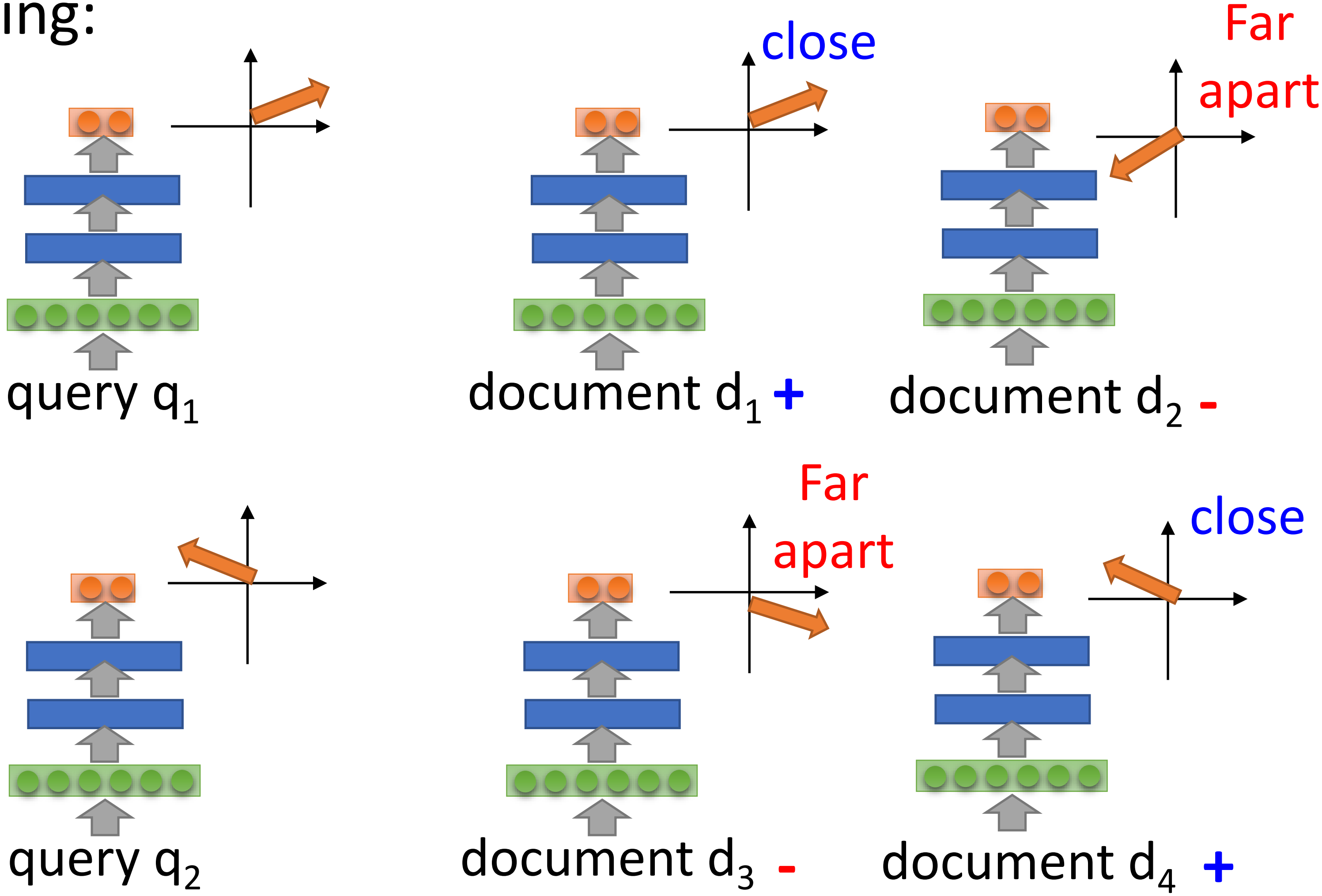
DSSM

Click-through data: $q_1 \longrightarrow d_1$: + d_2 : -

$q_2 \longrightarrow d_3$: - d_4 : +

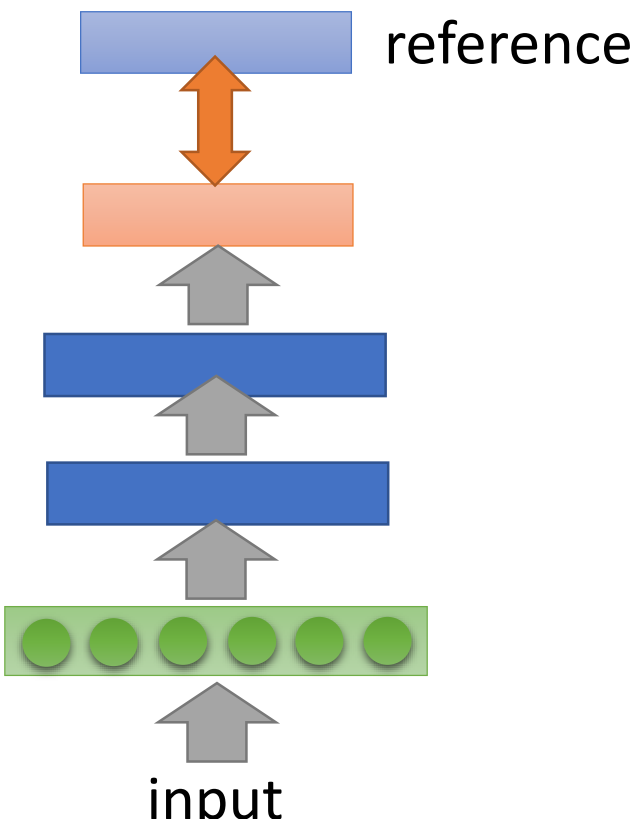
……

Training:

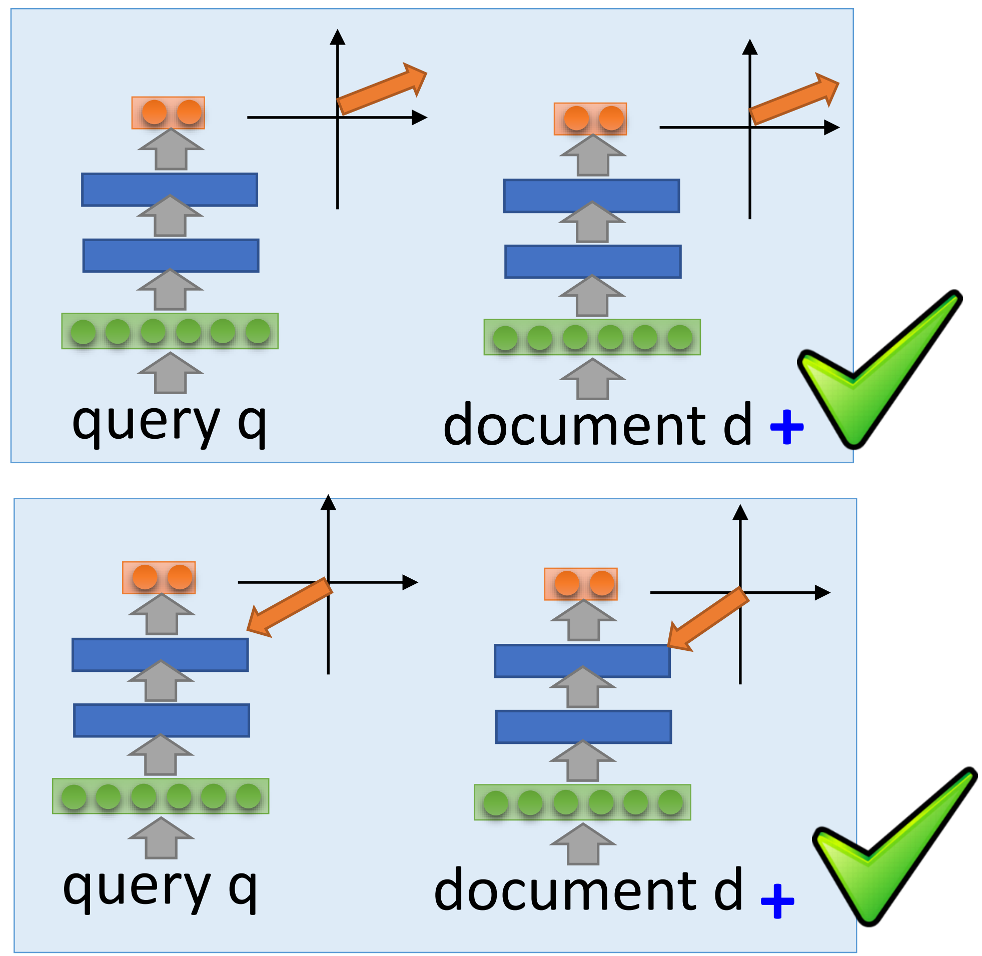


DSSM v.s. Typical DNN

Typical DNN



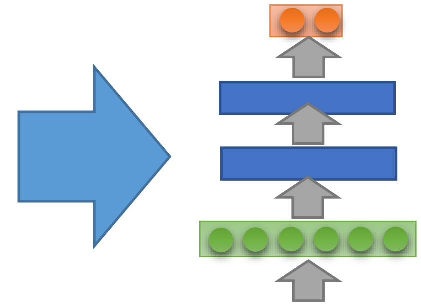
DSSM



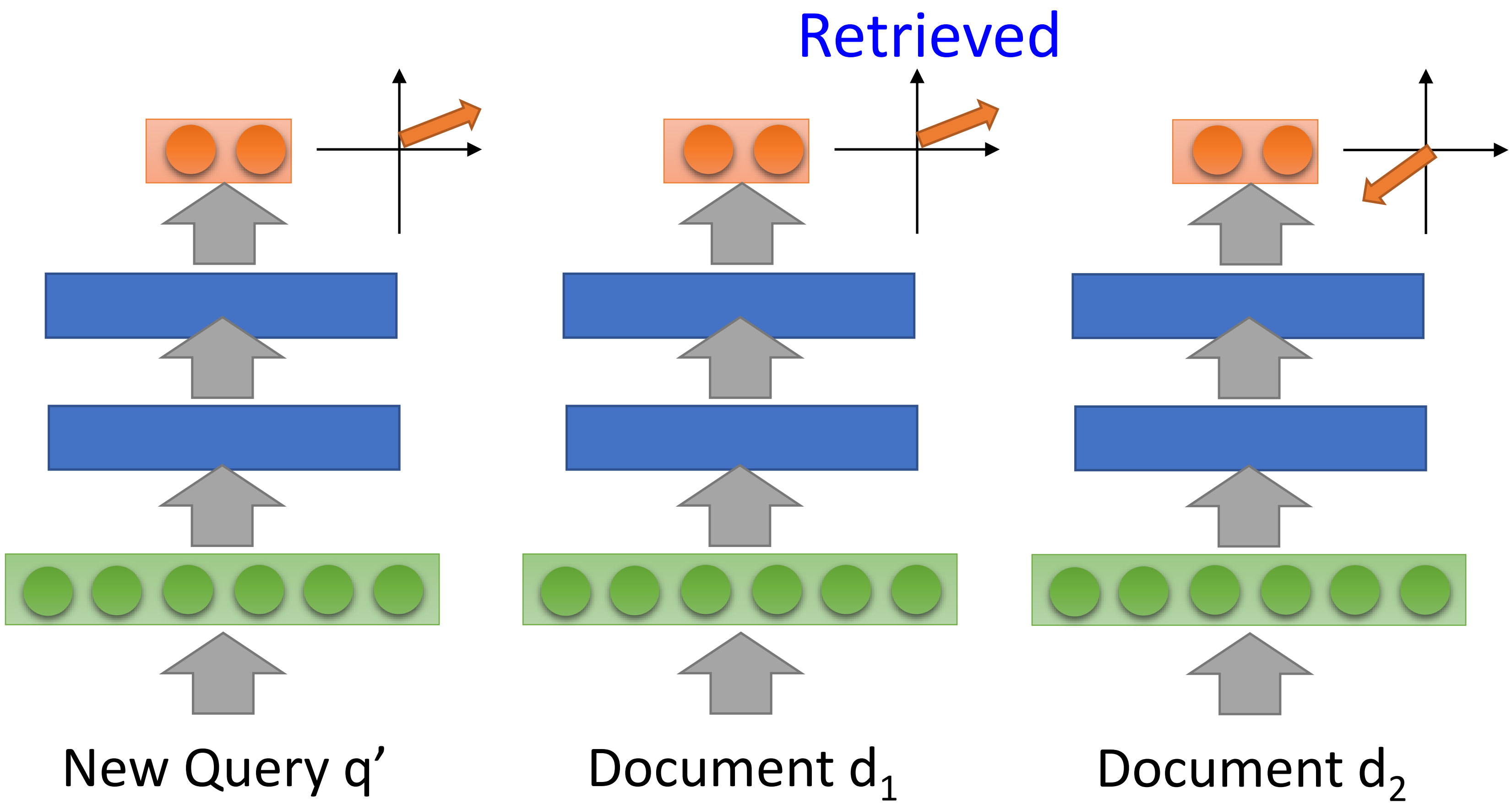
Click-through data: $q_1 \longrightarrow d_1$: + d_2 : -

$q_2 \longrightarrow d_3$: - d_4 : +

……

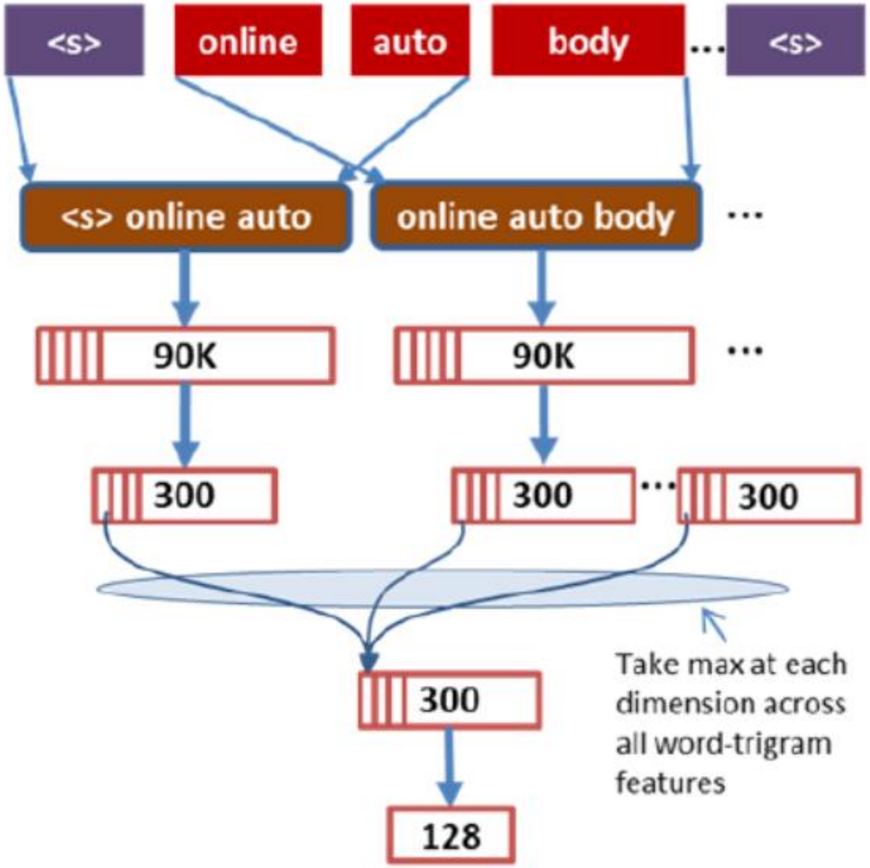


- How to do retrieval?



More …

- **Convolutional DSSM**: http://www.iro.umontreal.ca/~lisa/pointeurs/ir0895-he-2.pdf



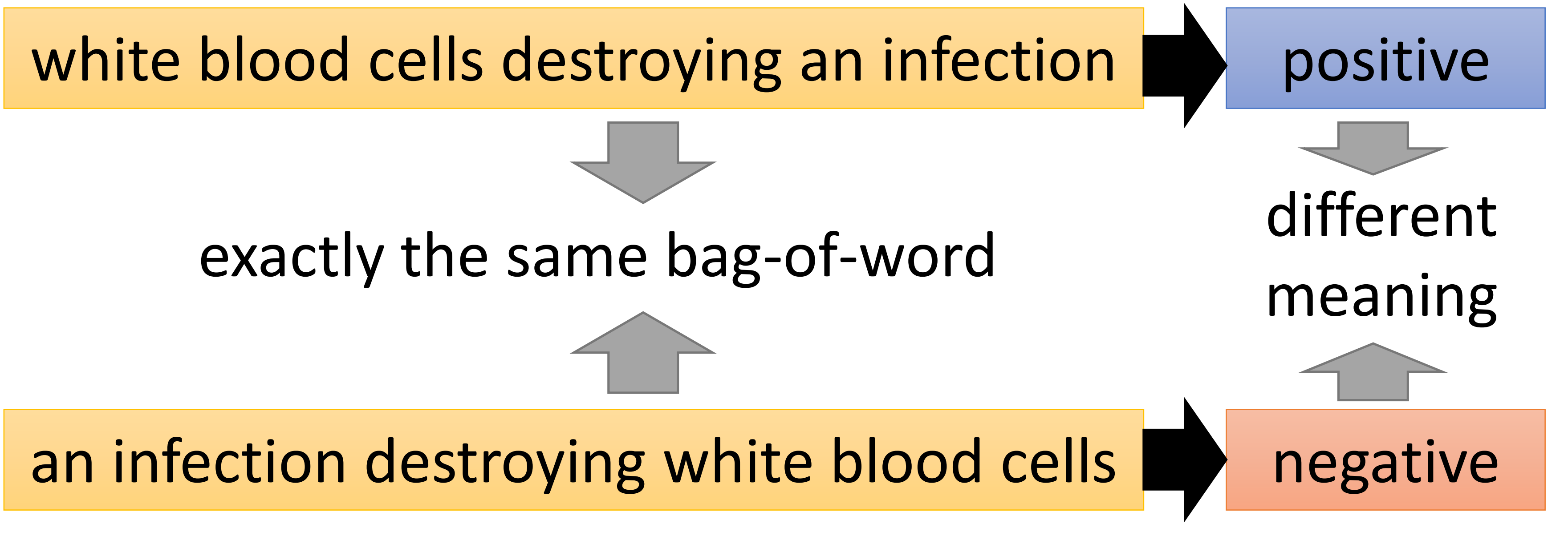
Meaning of Word Sequence - Outline

Deep Structured Semantic Model (DSSM)	• Application: Information Retrieval (IR)
Recursive Deep Model	• Application: Sentiment Analysis
Paragraph Vector	• Unsupervised

Reference: http://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf

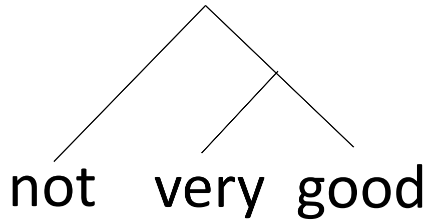
Recursive Deep Model

- When understanding the meaning of a word sequence, the order of the words can not be ignore.



Recursive Deep Model

syntactic structure



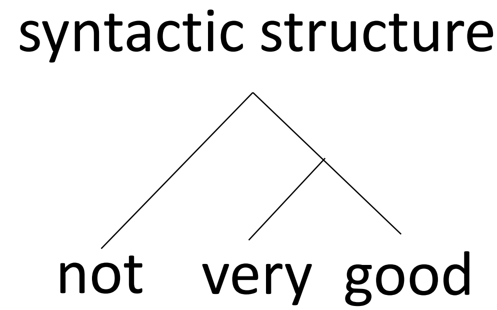
How to do it is out of the scope

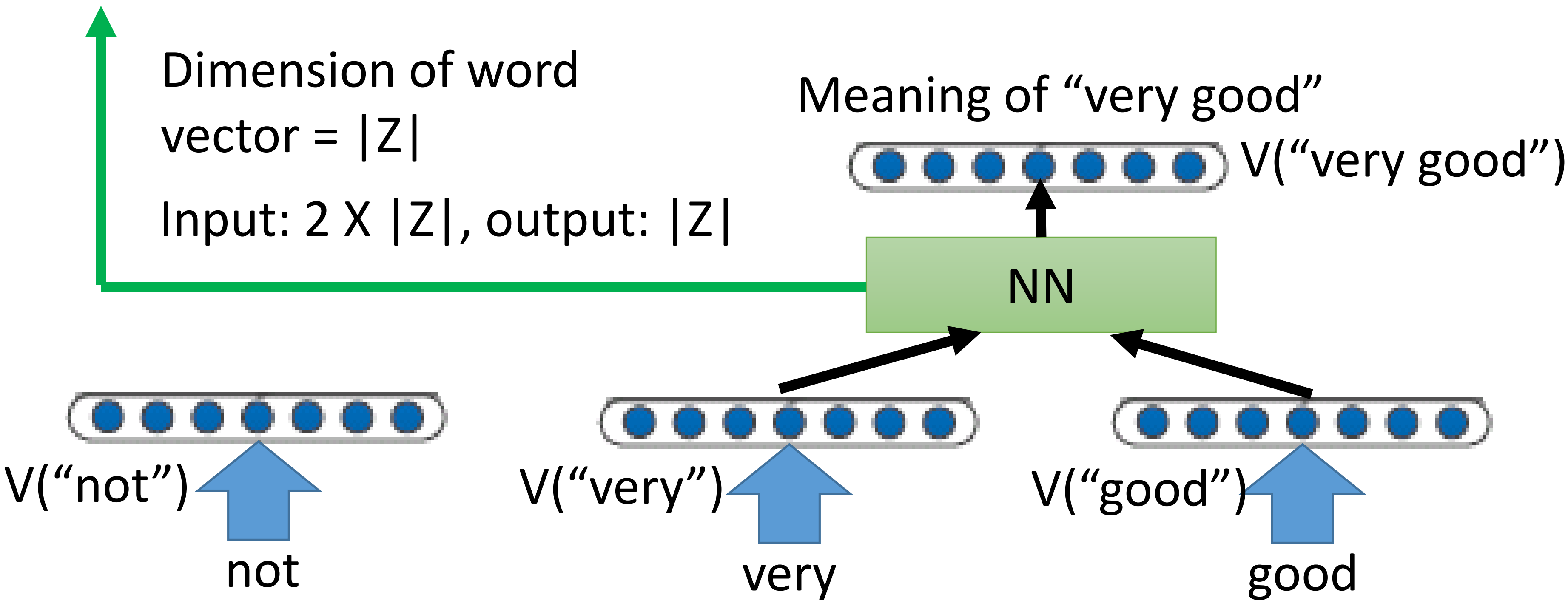
word sequence:

not very good

Recursive Deep Model

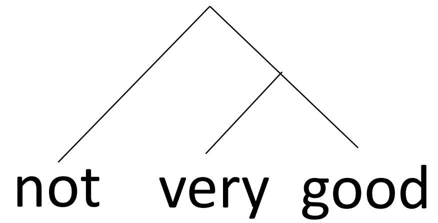
By composing the two meaning, what should the meaning be.





Recursive Deep Model

syntactic structure

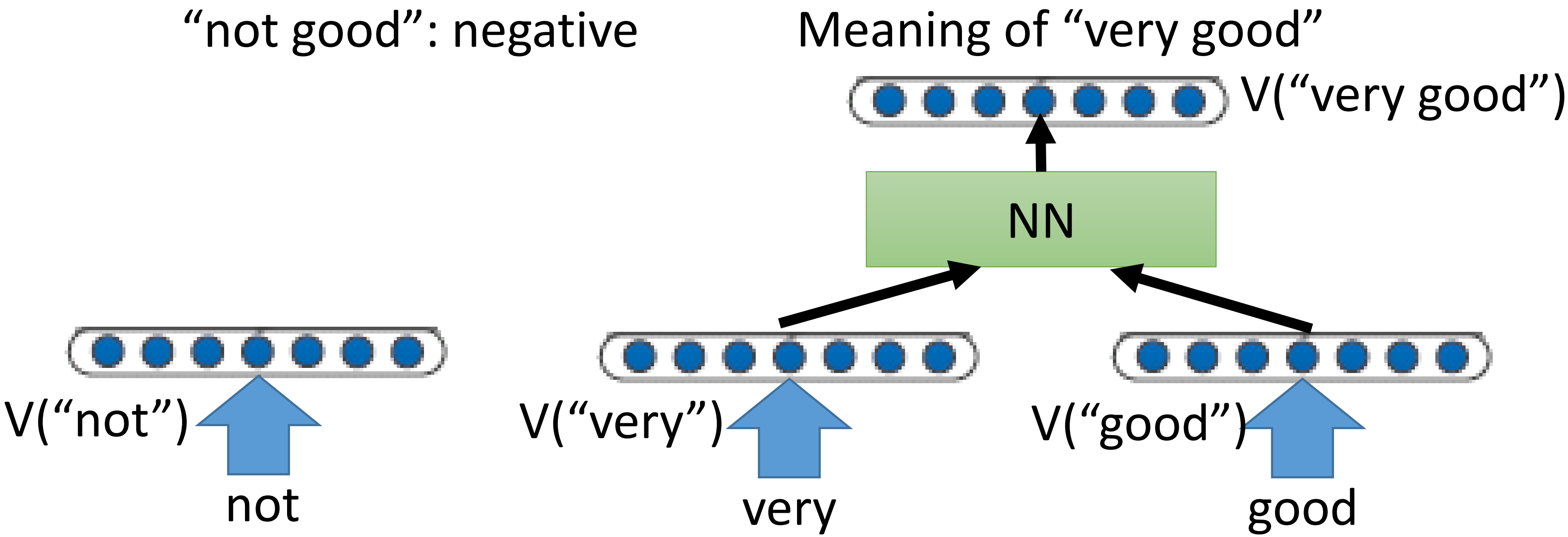


$V(w_A w_B) \neq V(w_A) + V(w_B)$

"not": neutral

"good": positive

"not good": negative



Recursive Deep Model

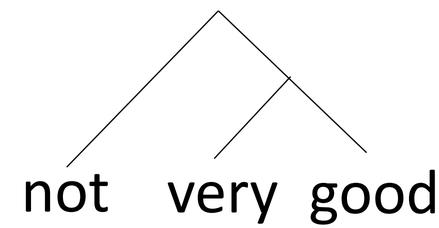
$V(w_A\, w_B) \neq V(w_A) + V(w_B)$

"棒": positive

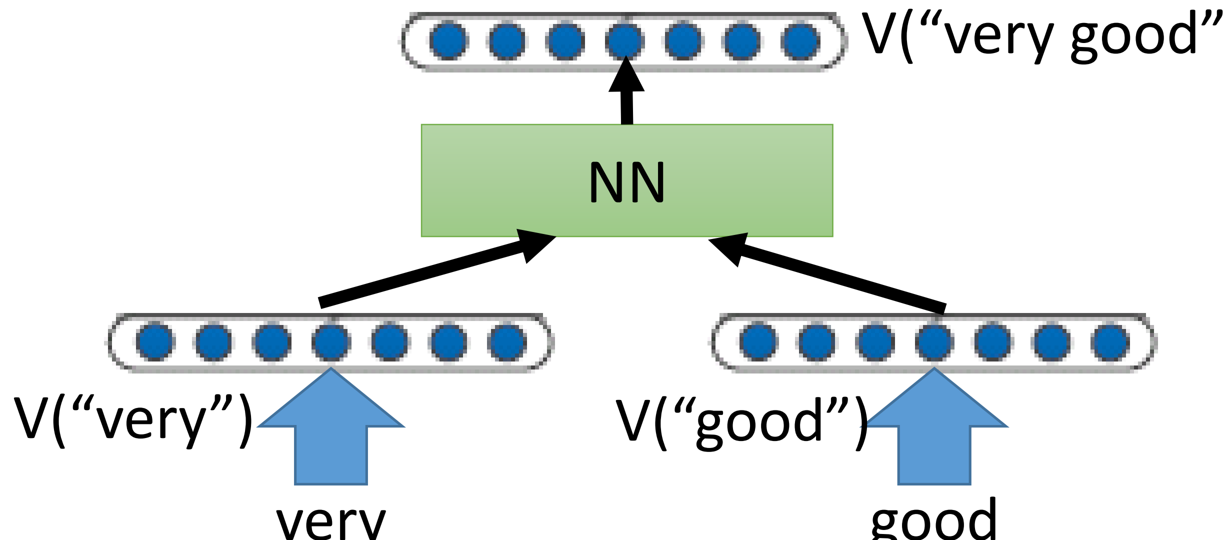
"好棒": positive

"好棒棒": negative

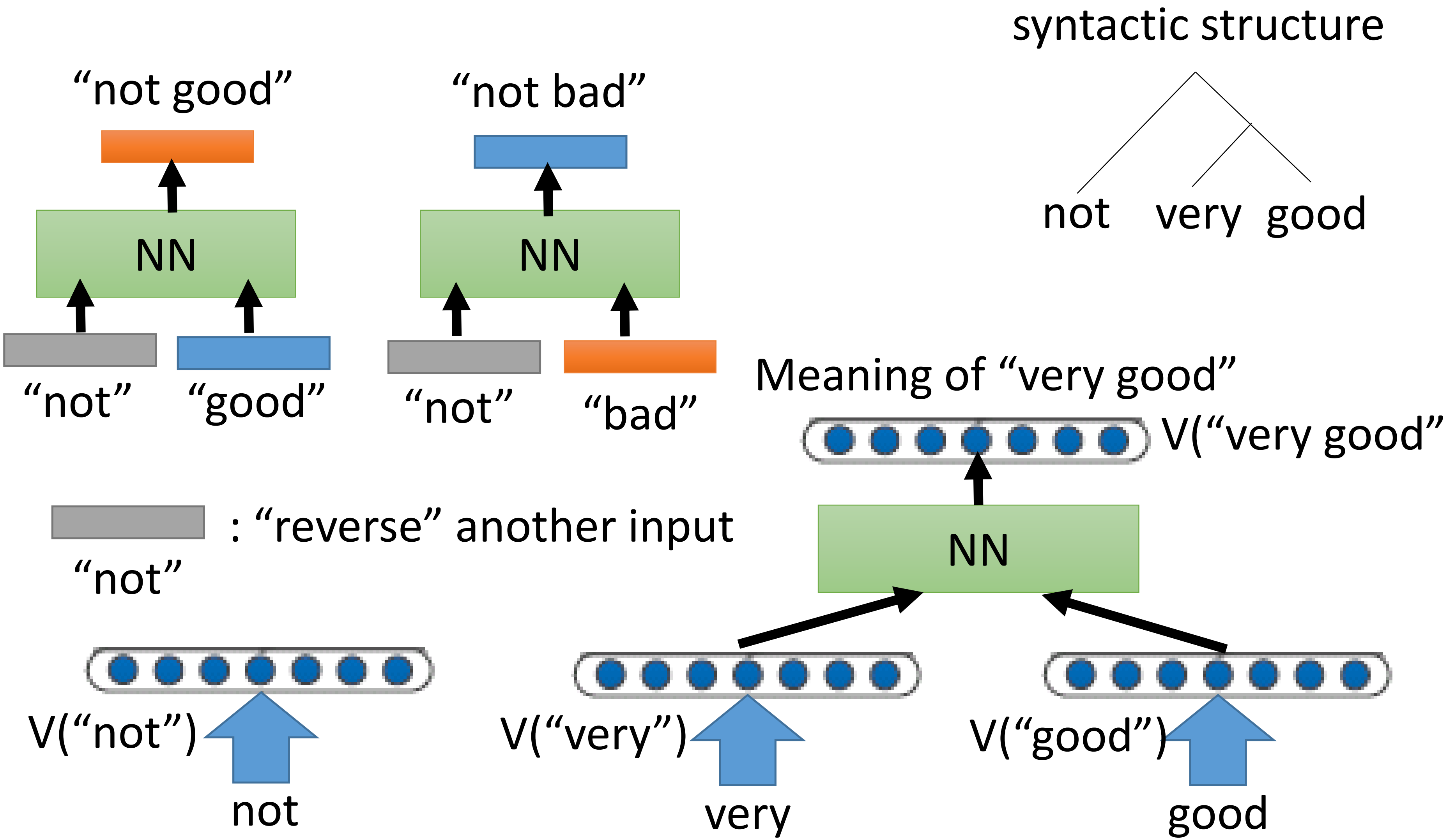
syntactic structure



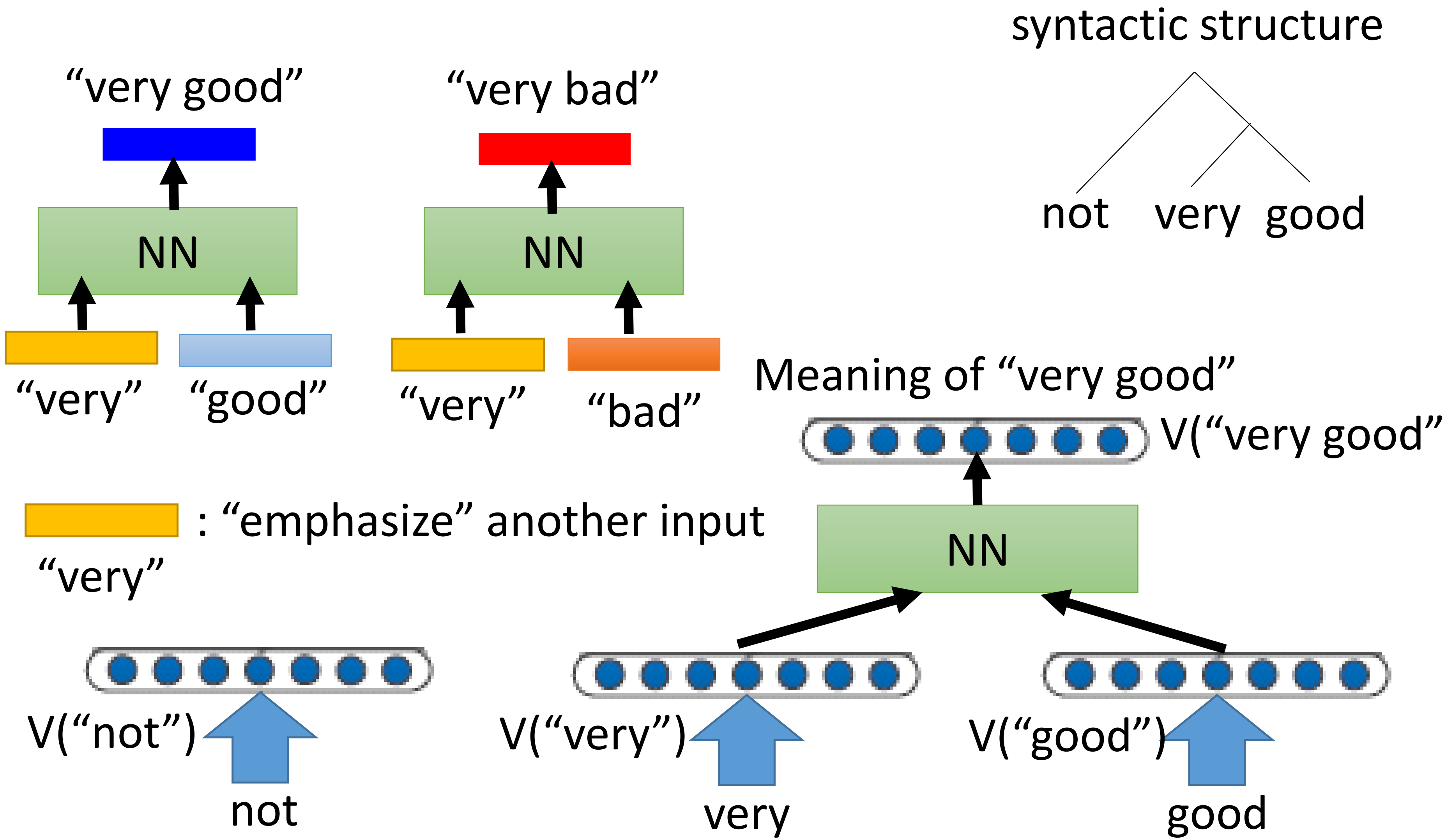
Meaning of "very good"



Recursive Deep Model



Recursive Deep Model

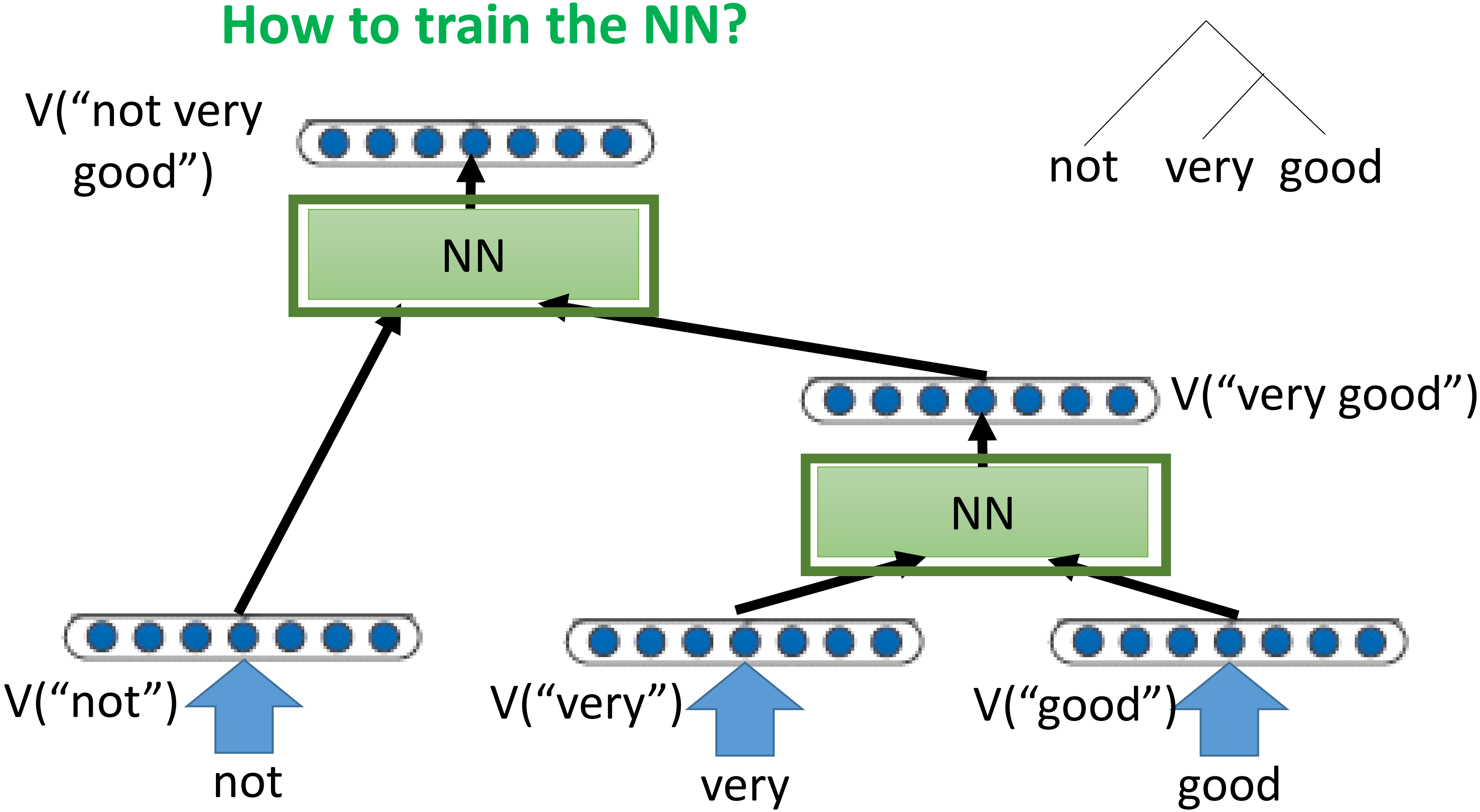


The word order is considered.

The representation of the sequence will change if the order of the words are changed

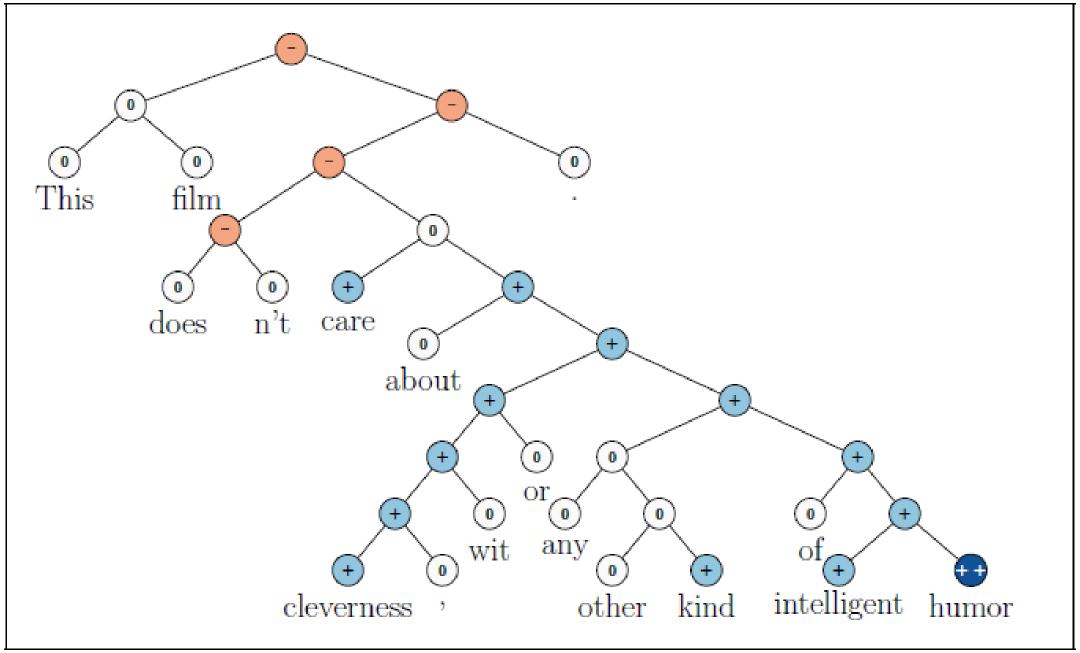
How to train the NN?

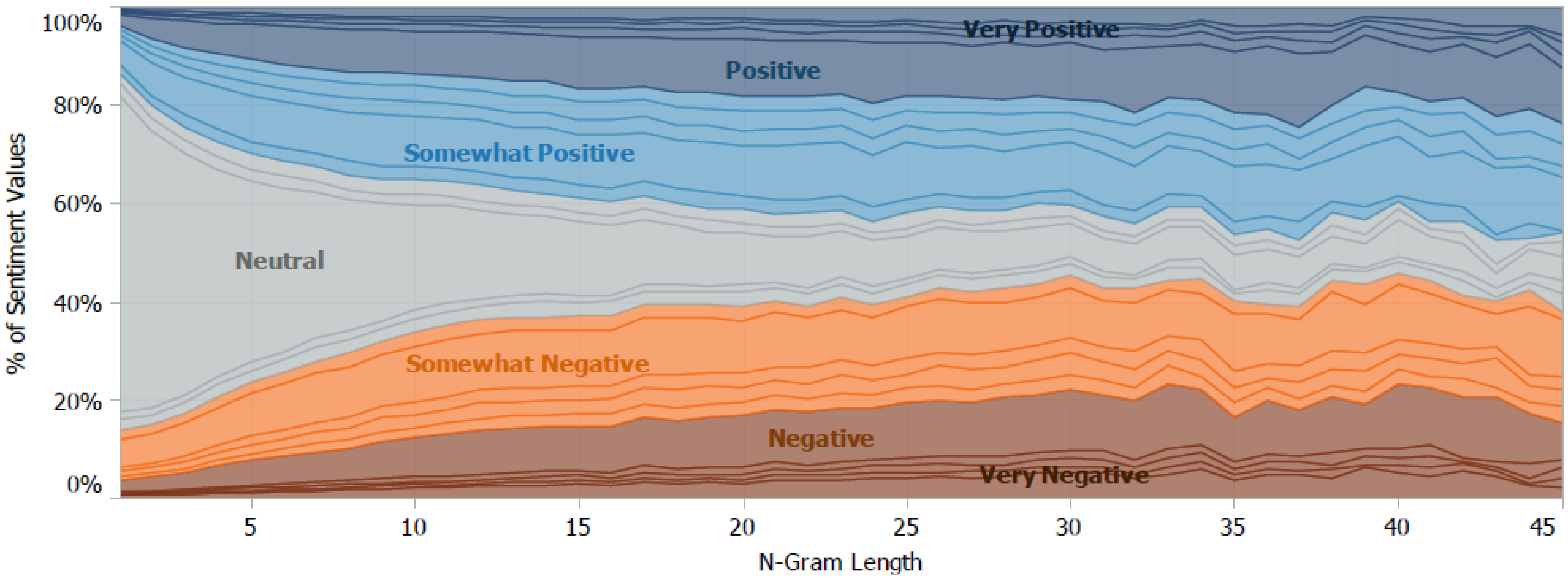
syntactic structure

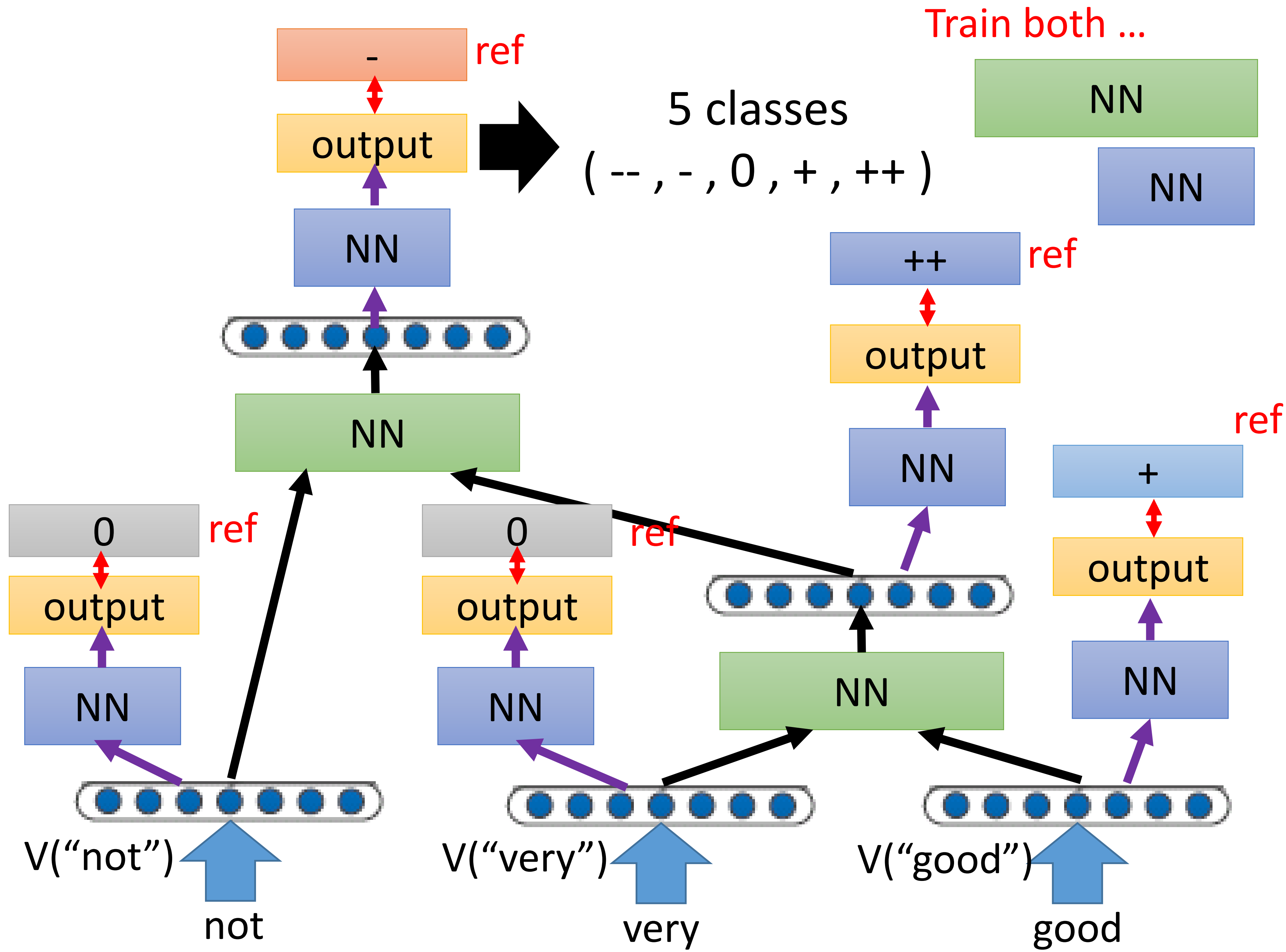


Training Data

5-class sentiment classification
(-- , - , 0 , + , ++)







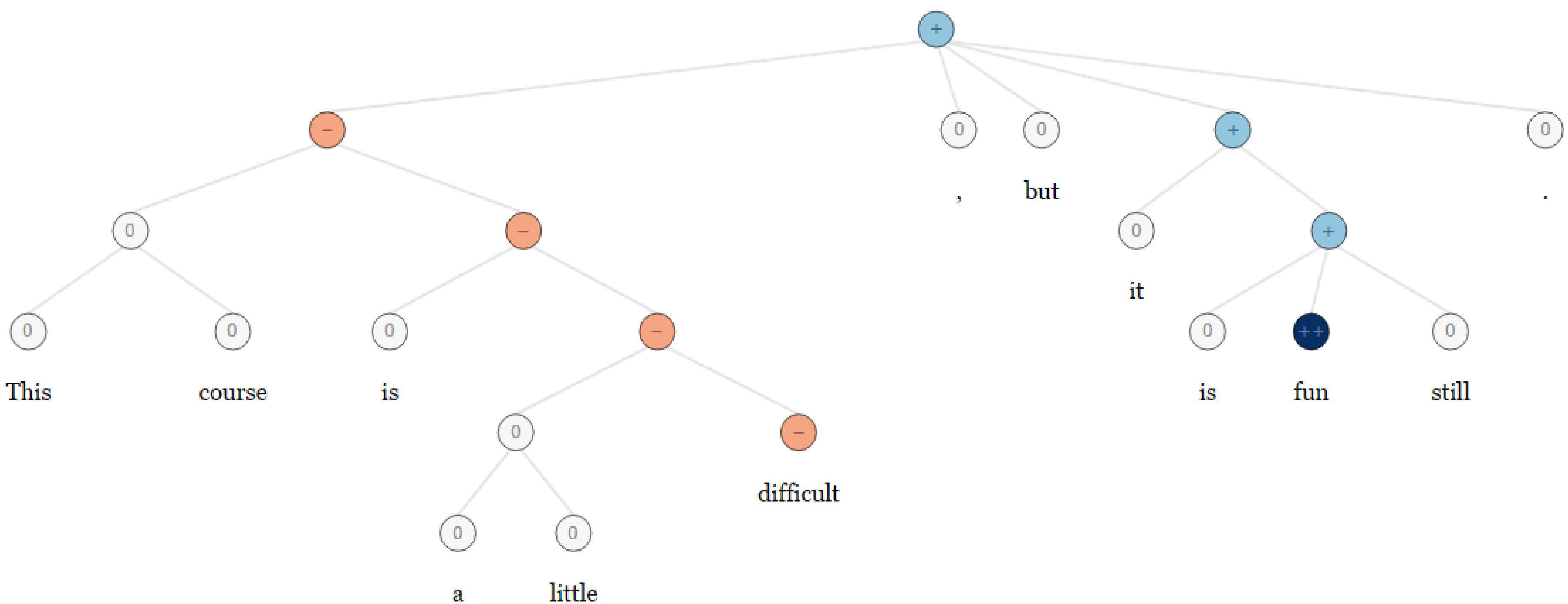
Train both ...
NN
NN
-
ref
output
5 classes
(-- , - , 0 , + , ++)
NN
NN
++
ref
output
NN
ref
+
0
ref
0
ref
output
output
output
NN
NN
NN
NN
V("not")
not
V("very")
very
V("good")
good

More ...

5-class sentiment classification
(-- , - , 0 , + , ++)

- Demo
  - http://nlp.stanford.edu:8080/sentiment/rntnDemo.html

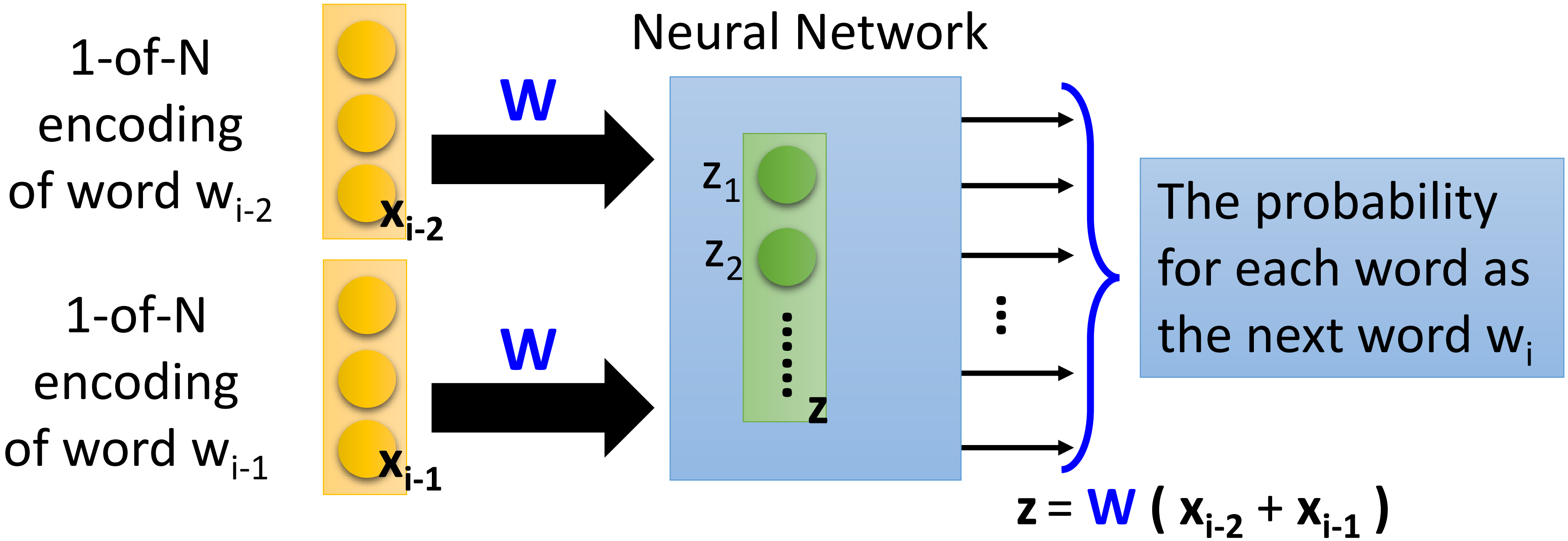
This course is a little difficult, but it is fun still.



Meaning of Word Sequence - Outline

Deep Structured Semantic Model (DSSM)	• Application: Information Retrieval (IR)
Recursive Deep Model	• Application: Sentiment Analysis
Paragraph Vector	• Unsupervised

Reference : http://cs.stanford.edu/~quocle/paragraph_vector.pdf



Paragraph d_1: (The paragraph is related to "The lord of the ring")

…… 魔君 (w_{i-2}) 名叫 (w_{i-1}) 索倫 (Sauron) (w_i) ……

$z = W (x_{i-2} + x_{i-1})$

Paragraph d_2: (The paragraph is related to "仙五")

…… 魔君 (w_{i-2}) 名叫 (w_{i-1}) 姜世離 (w_i) ……

$z = W (x_{i-2} + x_{i-1})$

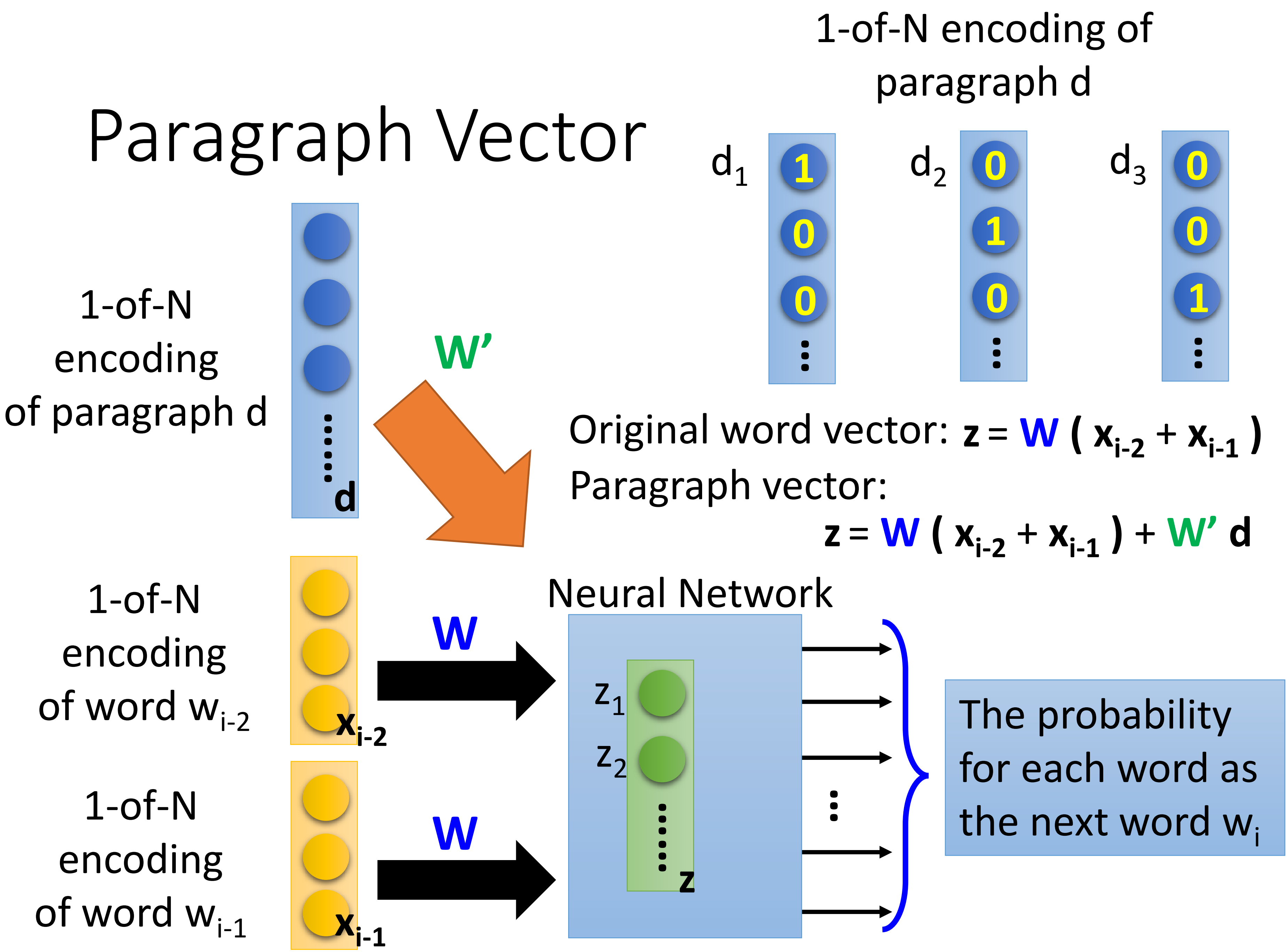
the same → Same output

Paragraph Vector

Original word vector: $\mathbf{z} = \mathbf{W} (\mathbf{x}_{i-2} + \mathbf{x}_{i-1})$

Paragraph vector:

$$\mathbf{z} = \mathbf{W} (\mathbf{x}_{i-2} + \mathbf{x}_{i-1}) + \mathbf{W'}\, \mathbf{d}$$



Paragraph Vector

Original word vector:

$$z = W(x_{i-2} + x_{i-1})$$

Paragraph vector:

$$z = W(x_{i-2} + x_{i-1}) + W' d$$

Then error of the prediction can be explained by the meaning of the paragraphs.

Paragraph d_1: (The paragraph is related to "The lord of the ring")

…… 魔君 名叫 索倫 (Sauron) ……

w_{i-2} w_{i-1} w_i

$$z = W(x_{i-2} + x_{i-1}) + W' d_1$$

different

Paragraph d_2: (The document is related to "仙五")

…… 魔君 名叫 姜世離 ……

w_{i-2} w_{i-1} w_i

$$z = W(x_{i-2} + x_{i-1}) + W' d_2$$

Paragraph vector of d: $V(d) = W' d$ → Meaning of the paragraph

Meaning of Word Sequence - Summary

Deep Structured Semantic Model (DSSM)	• Application: Information Retrieval (IR)
Recursive Deep Model	• Application: Sentiment Analysis
Paragraph Vector	• Unsupervised

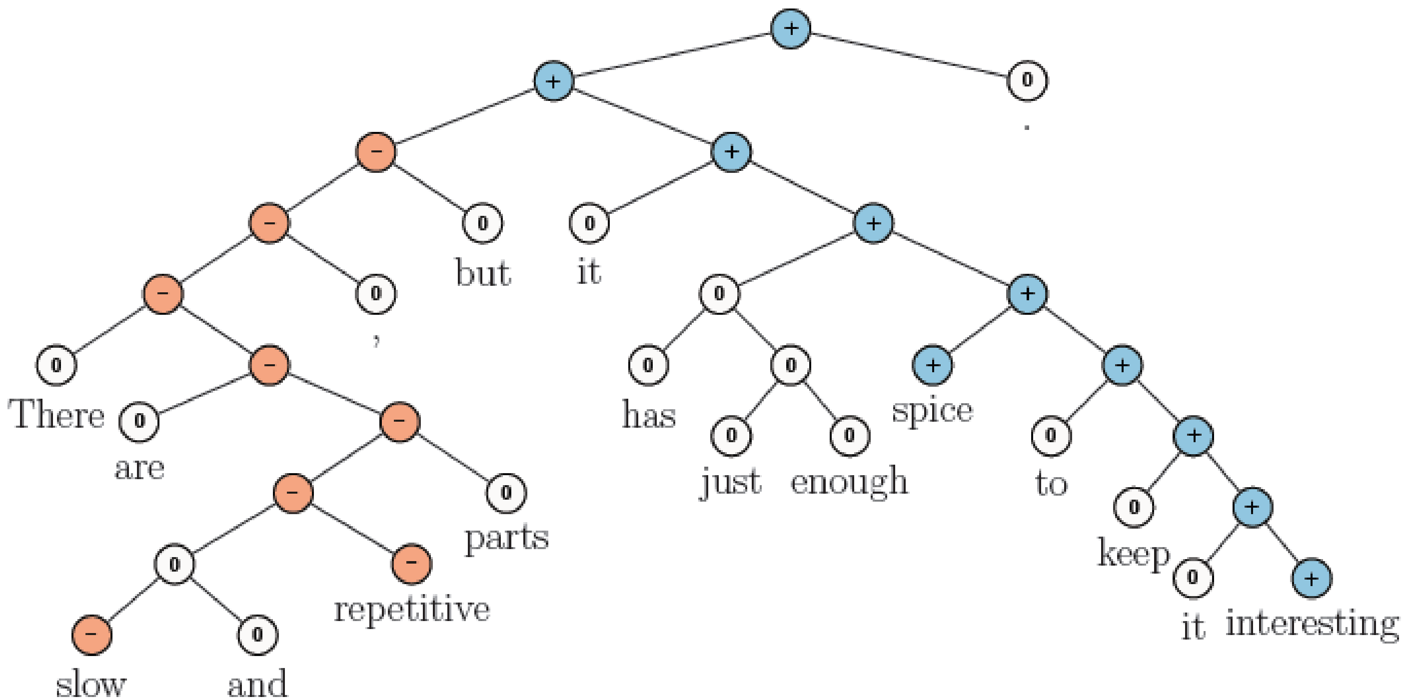

Thank You

Appendix

Chinese Room

https://www.youtube.com/watch?feature=player_embedded&v=0F3-j-GQcts

Demo in the paper



More ……

- The paragraph vector can also be used in retrieval
  - Demo: http://www.logos.t.u-tokyo.ac.jp/~hassy/implementations/paragraph_vector/
- Toolkit: https://github.com/klb3713/sentence2vec

Word classes

- One of the most successful NLP concepts in practice
- Similar words should share parameter estimation, which leads to generalization
- Example:

$$Class_1 = (yellow, green, blue, red)$$
$$Class_2 = (Italy, Germany, France, Spain)$$

- Usually, each vocabulary word is mapped to a single class (similar words share the same class)

Word classes

- There are many ways how to compute the classes – usually, it is assumed that similar words appear in similar contexts

- Instead of using just counts of words, we can use also counts of classes, which leads to generalization (better performance on novel data)

Class-based n-gram models of natural language (Brown, 1992)