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 0]
- bag = [0 1 0 0 0 0]
- cat = [0 0 1 0 0 0]
- dog = [0 0 0 1 0 0]
- elephant = [0 0 0 0 1 0]

**Word Embedding**

**Word Class**

- class 1: dog, cat, bird
- Class 2: ran, jumped, walk
- Class 3: flower, tree, apple
Word Embedding

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

蔡英文、馬英九 are something very similar

You shall know a word by the company it keeps
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

\[ V(w_i) \cdot V(w_j) \quad \text{Inner product} \quad N_{i,j} \quad \text{Number of times } w_i \text{ and } w_j \text{ in the same document} \]

• **Prediction 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
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

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

推 AO56789: 我同學才扯好不好，他有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

著名簽名檔 (出處不詳)
Prediction-based – Language Modeling

\[ P(\text{“wreck a nice beach”}) \]
\[ = P(\text{wreck} | \text{START}) P(\text{a} | \text{wreck}) P(\text{nice} | \text{a}) P(\text{beach} | \text{nice}) \]

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

1-of-N encoding of “START”
1-of-N encoding of “wreck”
1-of-N encoding of “a”
1-of-N encoding of “nice”
Prediction-based

The probability for each word as the next word $w_i$

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

- 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

You shall know a word by the company it keeps

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

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

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

The length of $x_{i-1}$ and $x_{i-2}$ are both $|V|$. The length of $z$ is $|Z|$. 

$z = W_1 x_{i-2} + W_2 x_{i-1}$

The weight matrix $W_1$ and $W_2$ are both $|Z| \times |X| \times |V|$ matrices.

$W_1 = W_2 = W \quad \Rightarrow \quad z = W ( x_{i-2} + x_{i-1} )$
Prediction-based
– Sharing Parameters

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

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

The weights with the same color should be the same.

Or, one word would have two word vectors.

The probability for each word as the next word $w_i$
Prediction-based – Various Architectures

• Continuous bag of word (CBOW) model

• Skip-gram

predicting the word given its context

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 \( V(w) \)
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)
Multi-lingual Embedding

Document Embedding

• 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
Semantic Embedding

Beyond Bag of Word

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

white blood cells destroying an infection

exactly the same bag-of-word

an infection destroying white blood cells

positive

different meaning

negative
Beyond Bag of Word

- **Paragraph Vector**: Le, Quoc, and Tomas Mikolov. "Distributed Representations of Sentences and Documents." ICML, 2014


Acknowledgement

• 感謝 John Chou 發現投影片上的錯字