Lec 1. Introduction and Word Vectors

- **Definition of 'meaning' in commonest linguistic way**

  : a pairing between a word (signifier or symbol)

  and signified thing (or an idea)

  => denotational semantics

  * semantics : 의미론

- **How do we have usable meaning in computer?**

  → WordNet : a thesaurus containing lists of synonyms and hypernyms

  ↳ 문제점 : missing nuance / missing new meanings / subjective ··· etc

  * thesaurus : 유의어 사전
  synonym : 동의어, 유의어
  hypernym : 상의어

- **Representing words as discrete symbols**

  → one-hot vector

  ↳ 문제점 : there's no natural notion of similarity

  => learn to encode similarity in the vectors themselves

  => Distributional semantics : 자주 함께 or 가까이 등장하는 단어를 의미론적으로 가장 비슷한 것

  when word w appears in a text,

  its **context** is the set of words that appear nearby

  (within fixed size of window)

- **word vectors**

  = (word) embeddings, (neural) word representations

  벡터의 내적 (dot product) 으로 similarity 계산

- **Word2vec**

: framework for learning word vectors

every word in a fixed library is represented by a vector.

keep adjusting word vectors to maximise $P(O|C)$, where [O : context words / C : center word]

- ex) … problems turning into banking crises as …

$P(w_{t+2}|w_t)$, $P(w_{t+1}|w_t)$, $P(w_{t-1}|w_t)$, $P(w_{t-2}|w_t)$

into: center word at position t

window size = 2 일 때 center word 기준 앞·뒤 2개씩의 단어들만 확률 계산

center word를 옮겨가며 확률 계산

- Likelihood $= L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \ne 0}} P(w_{t+j}|w_t;\theta)$ where [θ : all variables to be optimized / $t = 1, \cdots T$: position / m : size of window]

- objective function $= J(\theta) = -\frac{1}{T}\log L(\theta)$ (cost or loss func.)

$$= -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m \le j \le m \\ j \ne 0}} \log P(w_{t+j}|w_t;\theta)$$

- goal : maximizing predictive accuracy

⇒ minimizing objective function

- $P(w_{t+j}|w_t;\theta)$ 계산 방법

: [w가 center word (c) 일 때의 벡터 : v_w / w가 context word (o) 일 때의 벡터 : u_w]

⇒ $P(O|C) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V}\exp(u_w^T v_c)}$ → 일종의 softmax function (내적값을 0~1 사이의 확률값으로 변환)

$$\frac{\partial}{\partial v_c} P(O|C) = \underbrace{\frac{\partial}{\partial v_c}\log\exp(u_o^T v_c)}_{= \frac{\partial}{\partial v_c} u_o^T v_c = u_o} - \underbrace{\frac{\partial}{\partial v_c}\log\sum_{w \in V}\exp(u_w^T v_c)}_{= \frac{1}{\sum_{w \in V}\exp(u_w^T v_c)} \cdot \frac{\partial}{\partial v_c}\sum_{x \in V}\exp(u_x^T v_c)}$$

$$= u_o - \frac{\sum_{x \in V}\exp(u_x^T v_c)\cdot u_x}{\sum_{w \in V}\exp(u_w^T v_c)} = u_o - \sum_{x \in V}\frac{\exp(u_x^T v_c)}{\sum_{w \in V}\exp(u_w^T v_c)} \cdot u_x \quad (\leftarrow P(x|c))$$

$$= u_o - \sum_{x \in V} P(x|c)\, u_x$$

⇒ observed − expected