

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 : 의미

- 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 : 동의어 사전
synonym : 동의어, 동의어
hypernym : 상위어

- 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 등장한 단어를 의미론적으로 가깝다고 보는 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

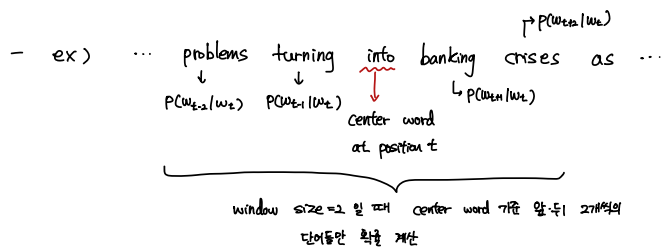
벡터의 내적 (dot product) 으 similarity 계산

• 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 $\begin{cases} o: \text{context words} \\ c: \text{center word} \end{cases}$



center word를 옮겨가며 확률 계산

- Likelihood = $L(\theta) = \prod_{t=1}^T \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} P(w_{t+j} | w_t; \theta)$ where $\begin{cases} \theta: \text{all variables to be optimized} \\ t=1, \dots, T: \text{position} \\ m: \text{size of window} \end{cases}$

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

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

- goal: maximizing predictive accuracy

\Rightarrow minimizing objective function

- $P(w_{t+j} | w_t; \theta)$ 계산 방법

$\begin{cases} w \text{가 center word } (c) \text{ 일 때의 벡터: } v_w \\ w \text{가 context word } (o) \text{ 일 때의 벡터: } u_w \end{cases}$

$\Rightarrow P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$ \rightarrow 일종의 softmax function
(내적값을 0~1 사이의 확률값으로 변환)

$$\begin{aligned} \frac{\partial}{\partial v_c} P(o|c) &= \frac{\frac{\partial}{\partial v_c} \log \exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} - \frac{\frac{\partial}{\partial v_c} \log \sum_{w \in V} \exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} \\ &= \frac{\frac{\partial}{\partial v_c} u_o^T v_c}{\sum_{w \in V} \exp(u_w^T v_c)} - \frac{\frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \cdot \frac{\partial}{\partial v_c} \sum_{w \in V} \exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} \\ &= \frac{u_o}{\sum_{w \in V} \exp(u_w^T v_c)} - \frac{\sum_{w \in V} \frac{\partial}{\partial v_c} \exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} \\ &= u_o - \sum_{w \in V} \frac{\exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} \cdot u_w \\ &= u_o - \sum_{w \in V} P(w|c) u_w \\ &\Rightarrow \text{observed} - \text{expected} \end{aligned}$$