

<https://www.edwith.org/deeplnlp/lecture/29208/>

CBoW (Continuous bag-of-words)

How to represent a sentence – CBoW

- Continuous bag-of-words
 - Ignore the order of the tokens: $(x_1, x_2, \dots, x_T) \rightarrow \{x_1, x_2, \dots, x_T\}$
 - Simply average the token vectors: $\frac{1}{T} \sum_{t=1}^T e_t$
 - Averaging is a differentiable operator.
 - Just one operator node in the DAG.
 - Generalizable to bag-of-n-grams
 - N-gram: a phrase of N tokens
 - *Think of how you would do!*
- Extremely effective in text classification [Iyyer et al., 2016; Cho, 2017; and many more]
 - For instance, if there are many positive words, the review is likely positive.
- In practice, use FastText [Bojanowski et al., 2017]

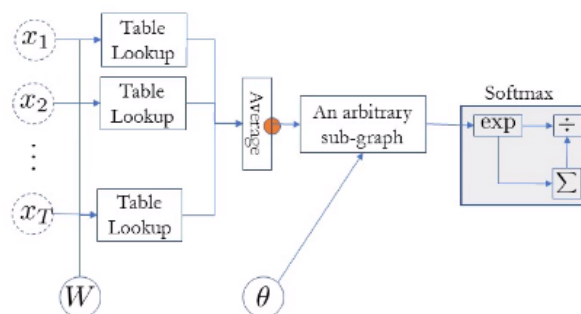
-토큰의 순서는 상관없다.

-문장에 대한 표현은 단어 벡터들을 평균시킨 벡터로 구한다.

-효과가 좋기 때문에 분석에서 제일 먼저 시도해야된다.

How to represent a sentence – CBoW

- Continuous bag-of-words based multi-class text classifier



- With this DAG, you use automatic backpropagation and stochastic gradient descent to train the classifier.

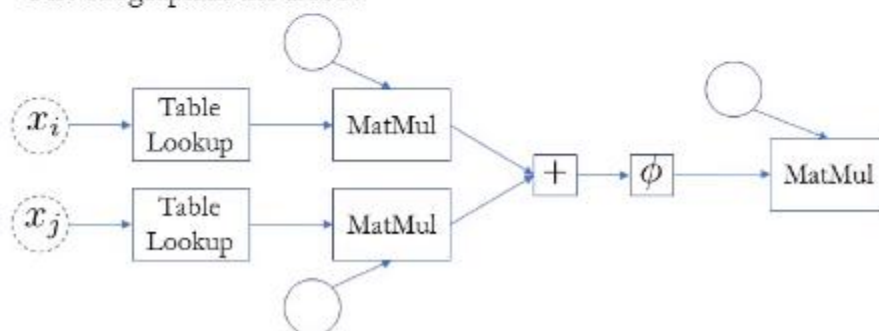
X는 토큰들(문장들).

공간상에서 문장이 가까우면 비슷한 의미, 멀리 떨어져 있으면 다른 의미.

text classification – 잘하는 애들끼리 모으고 그렇지 않은 애들끼리 모은다. 추후분석에 사용

How to represent a sentence – RN

- Relation Network [Santoro et al., 2017]: Skip Bigrams
 - Consider all possible pairs of tokens: $(x_i, x_j), \forall i \neq j$
 - Combine two token vectors with a neural network for each pair
 - $f(x_i, x_j) = W \phi(U_{\text{left}} e_i + U_{\text{right}} e_j)$
 - ϕ is a element-wise nonlinear function, such as tanh or ReLU ($\max(0, a)$)
 - One subgraph in the DAG.

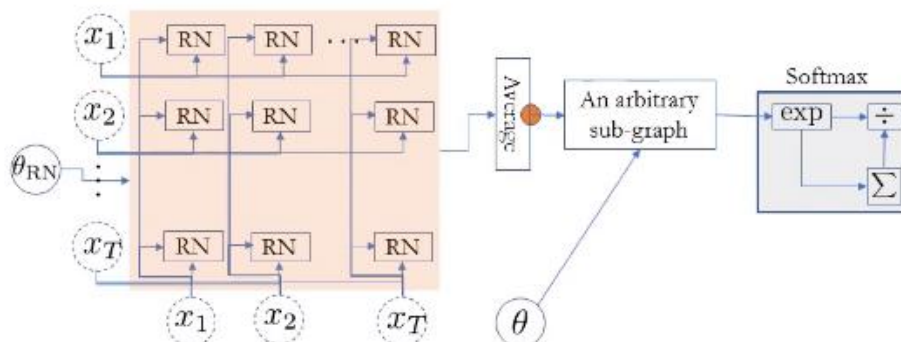


-단어 순서와 단어들 간의 관계 파악

-문장 안에 있는 모든 토큰 쌍(pair)을 본다. 모든 쌍에 대한 representative 찾기

How to represent a sentence – RN

- Relation Network: Skip Bigrams
 - Considers all possible pairs of tokens: $(x_i, x_j), \forall i \neq j$
 - Considers the pair-wise “relation”ship $f(x_i, x_j) = W \phi(U_{\text{left}} e_i + U_{\text{right}} e_j)$
 - Averages all these relationship vectors
- $$RN(X) = \frac{1}{2N(N-1)} \sum_{i=1}^{T-1} \sum_{j=i+1}^T f(x_i, x_j)$$



-모든 pair들의 interaction을 생각한다는 것이 앞에서 본 CBoW와 차별점

-장점: 여러 단어로 된 표현을 탐지 할 수 있다.

-단점: 모든 단어간의 관계를 보기 때문에 전혀 연관이 없는 단어도 보게 된다.

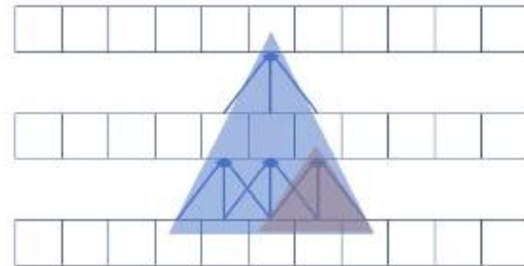
How to represent a sentence – CNN

- Convolutional Networks [Kim, 2014; Kalchbrenner et al., 2015]

- Captures k -grams hierarchically
- One 1-D convolutional layer: considers all k -grams

$$h_t = \phi \left(\sum_{\tau=-k/2}^{k/2} W_{\tau} e_{t+\tau} \right), \text{ resulting in } H = (h_1, h_2, \dots, h_T).$$

- Stack more than one convolutional layers: progressively-growing window
- Fits our intuition of how sentence is understood: **tokens**→**multi-word expressions**→**phrases**→**sentence**



-텍스트 데이터에서는 one-direction convolution

-k-gram을 계층적(hierachically)으로 보게 된다.(앞에k단어, 뒤에k단어)

-단어, 다중 단어 표현, 구절, 문장

-장점: 좁은 지역간 단어의 관계를 볼 수 있다.

-단점: 먼 거리의 단어간의 관계가 있을 경우 탐지할 수 없거나 더 많은 convolution층을 찾아야 된다.