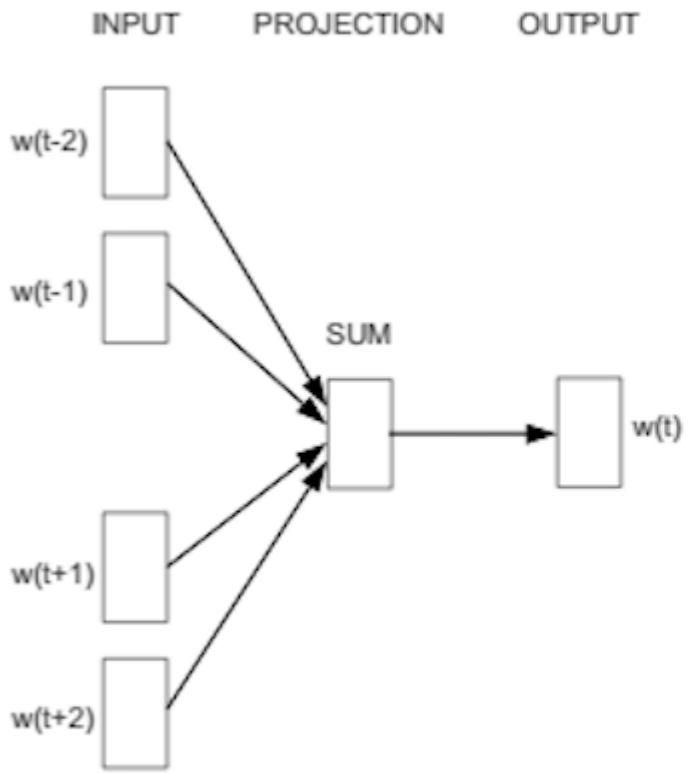
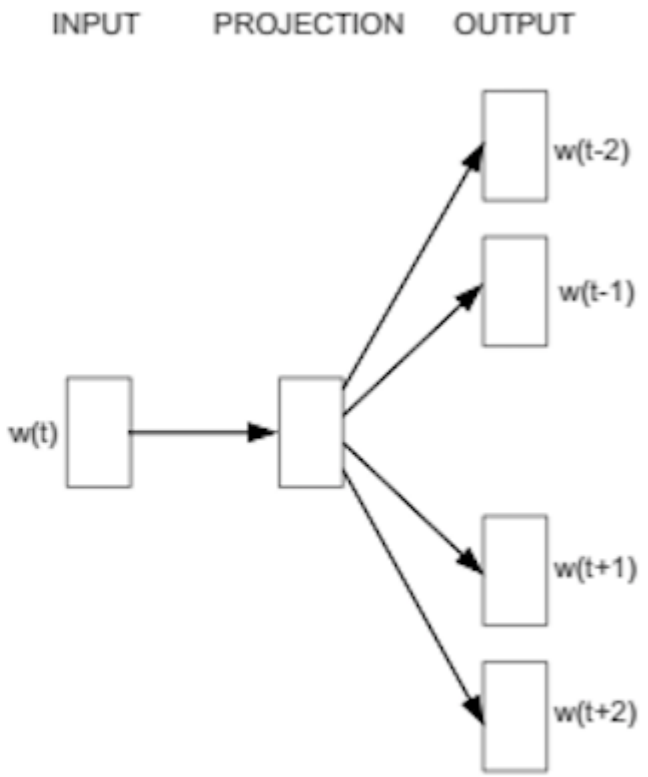
More Word Vectors

Minsik Park
(msp3887@korea.ac.kr)
Data Science & Business Analytics Lab

Additional word2vec



CBOW



Skip-gram

$$\min J = -u_c^t \hat{v} + \log \sum_{j=1}^{|V|} \exp(u_j^t \hat{v})$$

$$\min J = -\sum_{j=0, j\neq m}^{2m} u_{c-m+j}^t v_c + 2m \log \sum_{j=1}^{|V|} \exp(u_k^t v_c)$$

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Additional word2vec

$V \rightarrow \ln(V)$: complexity reduction

1. Hierarchical softmax

2. Negative Sampling

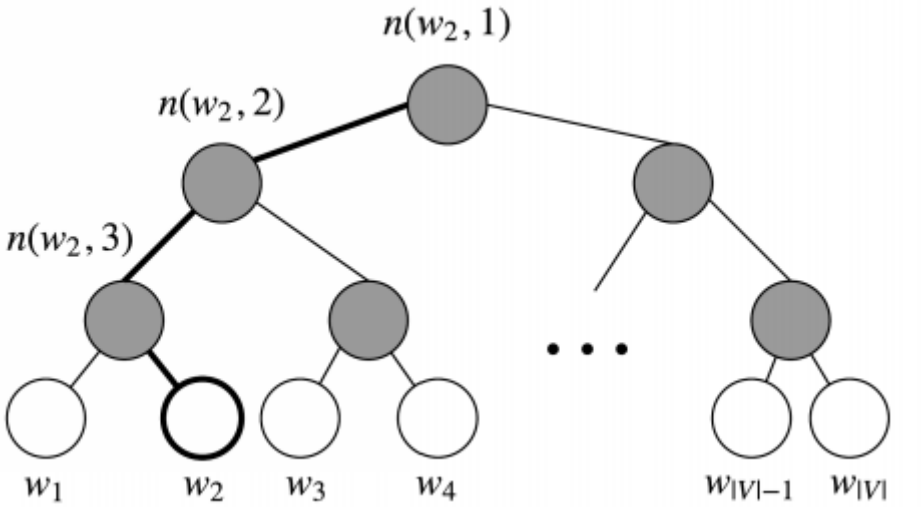
3. Subsampling Frequent words

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Additional word2vec

1. Hierarchical softmax

단어들을 leaves로 가지는 binary tree를 생성 후 해당하는 단어의 확률을 계산할 때 root 에서 부터 해당 leaf로 가는 path에 따라서 확률을 곱해나가는 식으로 해당 단어가 나올 최종적인 확률을 계산



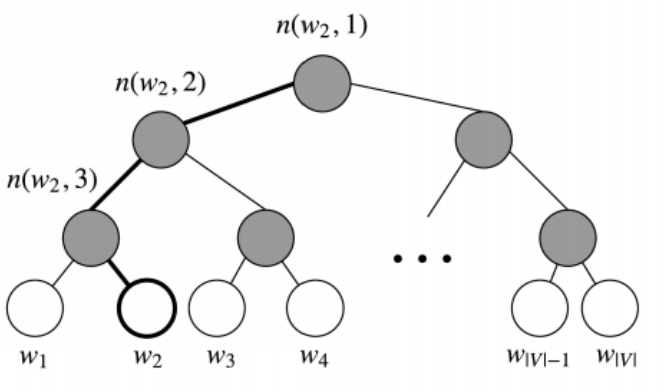
$$p(w|w_i) = \sum_{j=1}^{L(w)-1} \sigma([n(w, j+1) = ch(n(w, j))] \cdot V_{n(w,j)}^T v_{w_i})$$

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Additional word2vec

1. Hierarchical softmax

$$p(w|w_i) = \sum_{j=1}^{L(w)-1} \sigma([n(w,j+1) = ch(n(w,j))] \cdot V_{n(w,j)}^T v_{w_i})$$

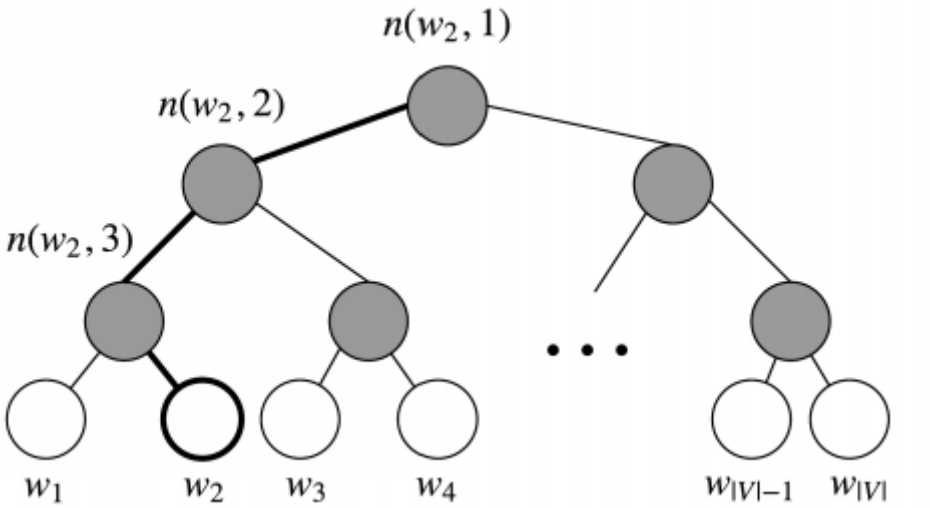


- L(w)는 w라는 leaf에 도달하기 까지의 path의 길이
- n(w,i) 는 root에서부터 w라는 leaf에 도달하는 path에서 만나는 i번째 노드를 의미
- ch(node) 는 node의 고정된 임의의 한 자식을 의미하며, 여기서는 단순하게 node의 왼쪽 자식의로 봐도 무방
- [[x]] 는 x가 true일 경우 1, false일 경우 -1을 반환하는 함수로 정의
- Hierarchial Softmax를 사용할 경우 기존 CBOW나 Skip-gram에 있던 W' matrix를 사용하지 않게 된다. 대신, V-1개의 internal node가 각각 길이 N짜리 weight vector를 가지게 된다. 이를 v_i' 라고 하고 학습에서 update 함
- root 에서 leaf까지의 거리의 평균은 O(lnV)가 되어 단어 벡터의 차원 수 감소

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Additional word2vec

1. Hierarchical softmax



$$p(w|w_i) = \sum_{j=1}^{L(w)-1} \sigma([n(w, j+1) = ch(n(w, j))] \cdot V_{n(w,j)}^T v_{w_i})$$

특정노드에서 왼쪽, 오른쪽 자식으로 갈 확률을 더하면 1

$$\sigma(v_n^T v_{w_i}) + \sigma(-v_n^T v_{w_i}) = 1$$

Hierarchical softmax를 사용하면 전체 확률에 대한 계산 없이 전체 합을 1로 만들어 줄 수 있음

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Additional word2vec

2. Negative Sampling

Softmax에서 너무 많은 단어들에 대해 계산을 하니,
여기서 몇 개(5~20개)만 샘플링해서 계산

실제 사용하는 단어들은 반드시 계산해야 되니 positive sample
나머지들은 negative sample

$$J_t(\theta) = log\sigma(u_0^T v_c) + \sum_{i=1}^{k} E_{i\sim p(w)}[log\sigma(-u_i^T v_c)]$$

Positive Sample

Negative sample

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Additional word2vec

2. Negative Sampling

단어가 추출될 확률은 Unigram Distribution의 3/4승

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=0}^{n}(f(w_i)^{3/4})}$$

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Additional word2vec

3. Subsampling Frequent words

'the', 'a', 'in' 등 자주 등장하는 단어들을 확률적으로 제외하여
학습속도 & 성능 향상

단어 w의 등장 빈도를 f(w)라 할 때, 학습할 때 각 단어는

$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$ 의 확률로 제외

t는 빈도가 일정값 이상일 때 제외한다는 threshold 값으로 논문에서는 10^{-5}를 사용

$f(w_i)$는 Corpus 내의 Term Frequency

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Glove(Global Vectors for Word Representation)

- Based on matrix factorization method

- GloVe consists of a weighted least squares model that trains on global word-word co-occurrence counts and thus makes efficient use of statistics.

* **Co-occurrence Matrix**

- X : co-occurrence matrix($V \cdot V$ dimension)
- X_{ij} : frequency of word i co-occurring with word j
- $X_i = \sum_k^V X_{ik}$: total number of occurrences of word i in corpus
- $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$: the probability of j appearing in the context of word i
- $w \in R^d$: a word embedding of dimension d
- $\widetilde{w} \in R^d$: a context word embedding of dimension d

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.

Glove(Global Vectors for Word Representation)

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Ratios of co-occurrence probabilities can encode meaning

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.

Glove(Global Vectors for Word Representation)

(1) $F\left(w_i, w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}}$

(2) $F\left(w_i - w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}}$

(3) $F\left((w_i - w_j)^T \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}}$

(4) $F\left((w_i - w_j)^T \widetilde{w_k}\right) = \frac{F(w_i^T \widetilde{w_k})}{F(w_j^T \widetilde{w_k})}$

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.

Glove(Global Vectors for Word Representation)

$$F\left(w_i^T \widetilde{w_k} - w_j^T \widetilde{w_k}\right) = \frac{F(w_i^T \widetilde{w_k})}{F(w_j^T \widetilde{w_k})}$$

$$exp\left(w_i^T \widetilde{w_k} - w_j^T \widetilde{w_k}\right) = \frac{exp(w_i^T \widetilde{w_k})}{exp(w_j^T \widetilde{w_k})}$$

$$w_i^T \widetilde{w_k} = \log P_{ik} = \log X_{ik} - \log X_i$$

$$w_i^T \widetilde{w_k} = \log X_{ik} - b_i - \widetilde{b_k}$$

$$w_i^T \widetilde{w_k} + b_i + \widetilde{b_k} = \log X_{ik}$$

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.

Glove(Global Vectors for Word Representation)

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \widetilde{w_j} + b_i + \widetilde{b_j} - \log X_{ik})^2$$

Where $f(x) = \begin{cases} (\frac{x}{x_{max}})^{\alpha}, if\ x < x_{max} \\ 1, otherwise \end{cases}$

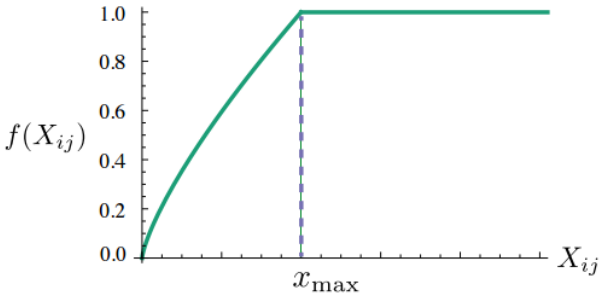
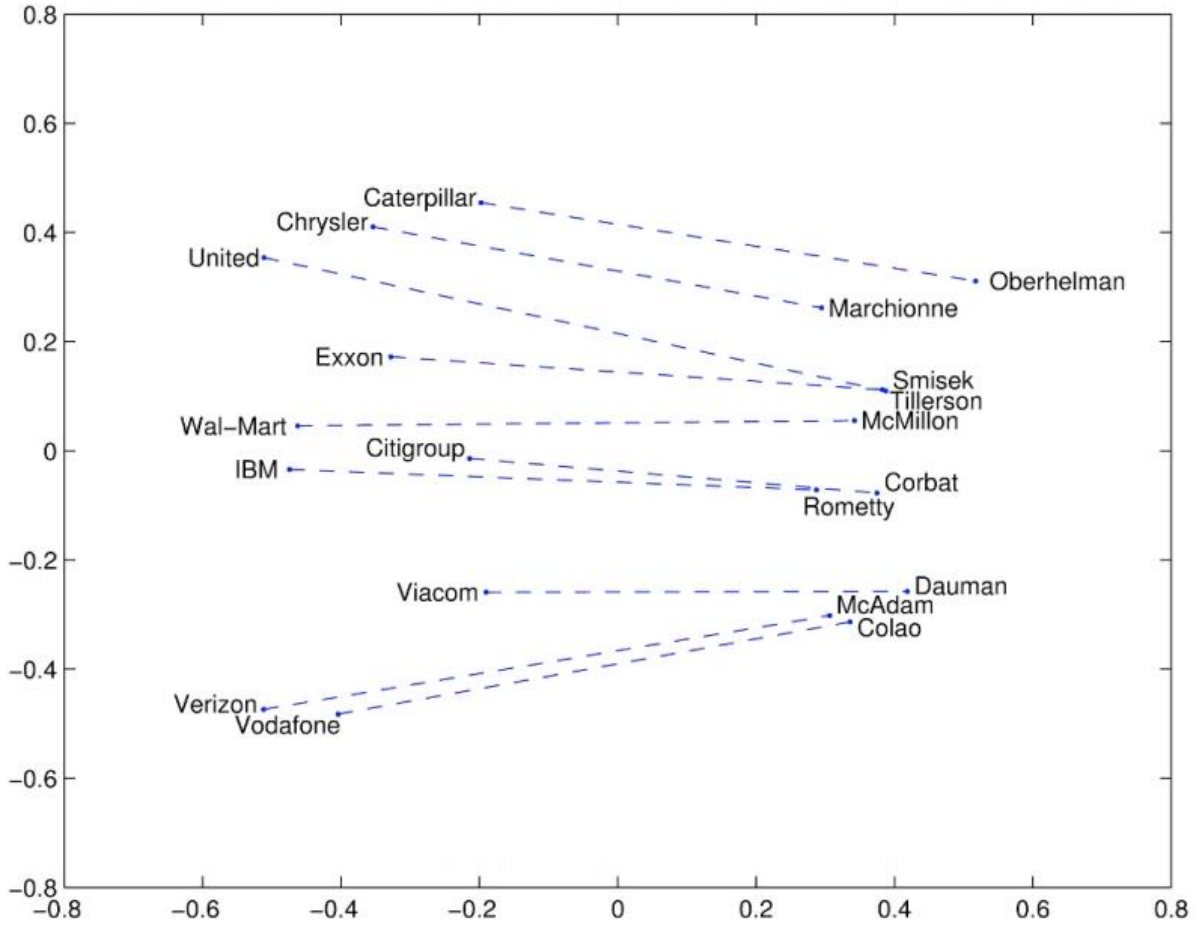


Figure 1: Weighting function f with $\alpha = 3/4$.

1. f(0) = 0
2. f(x) should be non-decreasing so that rare co-occurrences are not overweighted.
3. f(x) should be relatively small for large values of x, so the frequent co-occurences are not overweighted.

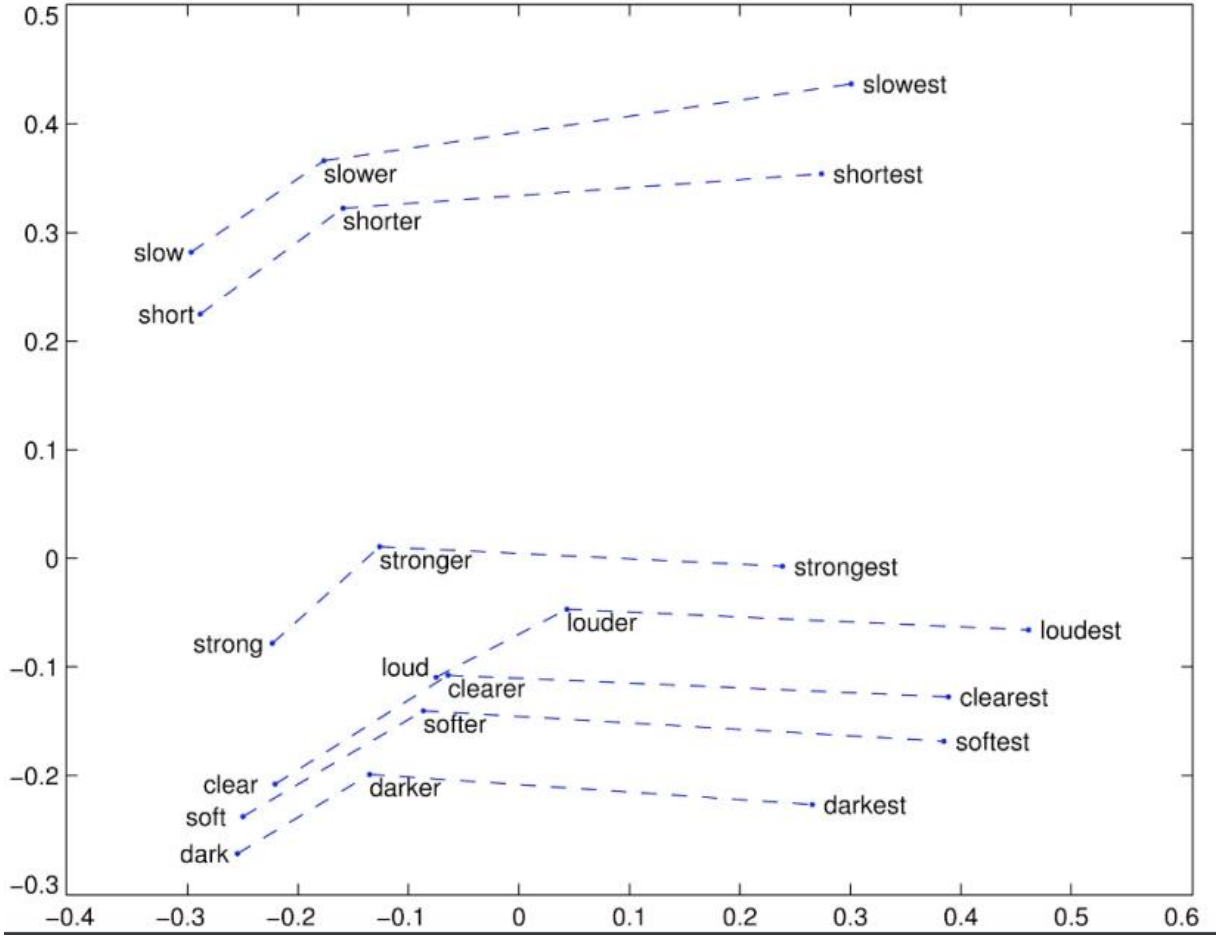
Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.

Glove visualization:Company-CEO



http://web.stanford.edu/class/cs224n/lecture_notes/cs224n-2017-notes2.pdf

Glove visualization:Superlatives



http://web.stanford.edu/class/cs224n/lecture_notes/cs224n-2017-notes2.pdf

Word Embedding Result(semantic)

Input	Result Produced
Chicago : Illinois : : Houston	Texas
Chicago : Illinois : : Philadelphia	Pennsylvania
Chicago : Illinois : : Phoenix	Arizona
Chicago : Illinois : : Dallas	Texas
Chicago : Illinois : : Jacksonville	Florida
Chicago : Illinois : : Indianapolis	Indiana
Chicago : Illinois : : Austin	Texas
Chicago : Illinois : : Detroit	Michigan
Chicago : Illinois : : Memphis	Tennessee
Chicago : Illinois : : Boston	Massachusetts

Input	Result Produced
Abuja : Nigeria : : Accra	Ghana
Abuja : Nigeria : : Algiers	Algeria
Abuja : Nigeria : : Amman	Jordan
Abuja : Nigeria : : Ankara	Turkey
Abuja : Nigeria : : Antananarivo	Madagascar
Abuja : Nigeria : : Apia	Samoa
Abuja : Nigeria : : Ashgabat	Turkmenistan
Abuja : Nigeria : : Asmara	Eritrea
Abuja : Nigeria : : Astana	Kazakhstan

http://web.stanford.edu/class/cs224n/lecture_notes/cs224n-2017-notes2.pdf

Word Embedding Result(syntactic)

Input	Result Produced
bad : worst : : big	biggest
bad : worst : : bright	brightest
bad : worst : : cold	coldest
bad : worst : : cool	coolest
bad : worst : : dark	darkest
bad : worst : : easy	easiest
bad : worst : : fast	fastest
bad : worst : : good	best
bad : worst : : great	greatest

Input	Result Produced
dancing : danced : : decreasing	decreased
dancing : danced : : describing	described
dancing : danced : : enhancing	enhanced
dancing : danced : : falling	fell
dancing : danced : : feeding	fed
dancing : danced : : flying	flew
dancing : danced : : generating	generated
dancing : danced : : going	went
dancing : danced : : hiding	hid
dancing : danced : : hitting	hit

http://web.stanford.edu/class/cs224n/lecture_notes/cs224n-2017-notes2.pdf

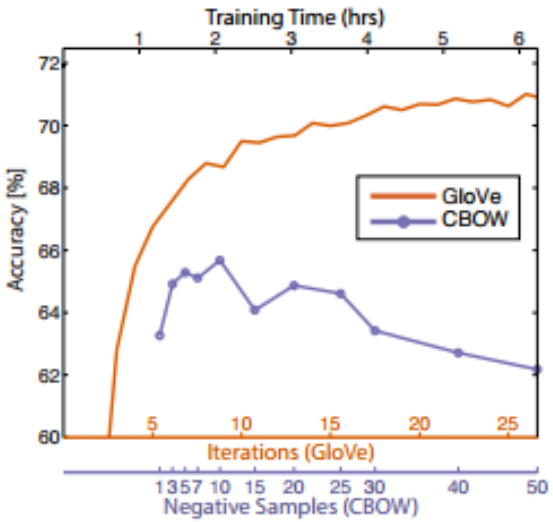
Word Embedding Result

Model	Dimension	Size	Semantics	Syntax	Total
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVE	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW	300	6B	63.6	67.4	65.7
SG	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	81.9	69.3	75.0

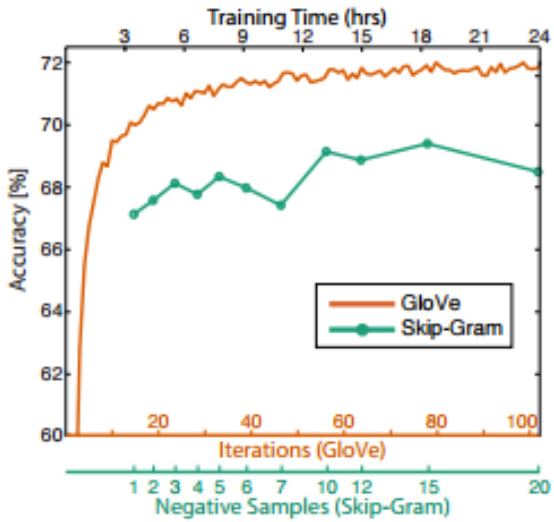
1. Performance is heavily dependent on the model used for word embedding
2. Performance increases with larger corpus sizes
3. Performance is lower for extremely low as well as for extremely high dimensional word vectors

http://web.stanford.edu/class/cs224n/lecture_notes/cs224n-2017-notes2.pdf

Comparison with Word2vec



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

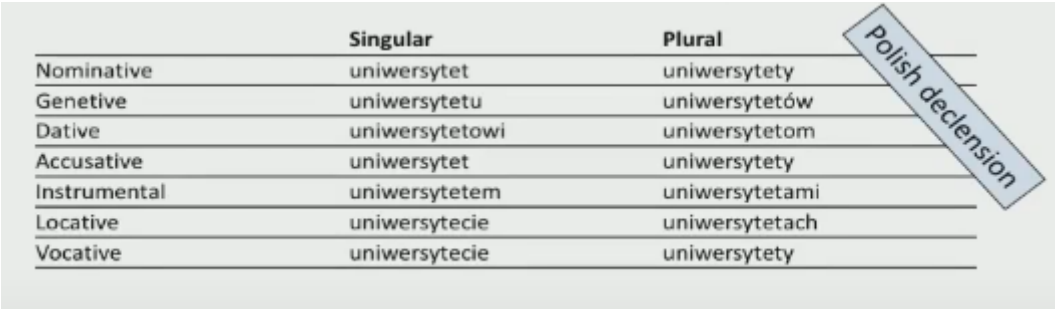
- 논문 전반적으로 GLoVe가 word2vec에 대해 확실한 성능 우세를 보인다고 주장
- 하지만 파라미터 세팅의 세밀함 등등 실험 평가에 대해서는 절대적으로 판단하기 어려움
- 이 논문에서도 word2vec의 cbow보다 skip-gram 성능이 전반적으로 높음을 확인할 수 있음

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.

FastText

- NNLM, Word2Vec, Glove ignore the morphology or words by assigning a distinct vector to each word

- Difficult to apply to morphologically rich languages with large vocabularies and many rare words (Turkish or Finnish)



	Singular	Plural
Nominative	uniwersytet	uniwersytety
Genetive	uniwersytetu	uniwersytetów
Dative	uniwersytetowi	uniwersytetom
Accusative	uniwersytet	uniwersytety
Instrumental	uniwersytetem	uniwersytetami
Locative	uniwersytecie	uniwersytetach
Vocative	uniwersytecie	uniwersytety

Polish declension

- FastText represent words as sum of its character n-grams

https://www.youtube.com/watch?v=CHcExDsDeHU

Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *arXiv preprint arXiv:1607.04606* (2016).

FastText

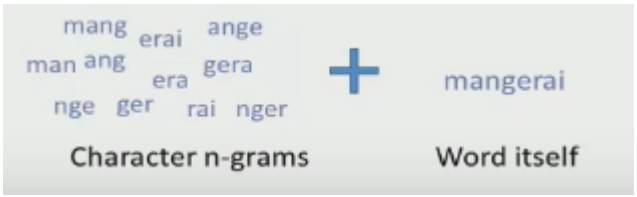
- Objective function

$$J_t(\theta) = \log \sigma(u_o^T v_c) + \sum_{i=1}^{k} E_{i \sim p(w)}[log\sigma(-u_i^T v_c)]$$

- Subword model

Define the set of n-grams appearing in w : $G_w \subset \{1,2, \dots G\}$

$$score(w, c) = \sum_{g \in G_w} z_g^T v_c$$



https://www.youtube.com/watch?v=CHcExDsDeHU

Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *arXiv preprint arXiv:1607.04606* (2016).

FastText

- Word analogies

Paris -> France

Warsaw -> ?

		sg	cbow	ours
CS	Semantic	25.7	27.6	27.5
	Syntactic	52.8	55.0	77.8
DE	Semantic	66.5	66.8	62.3
	Syntactic	44.5	45.0	56.4
EN	Semantic	78.5	78.2	77.8
	Syntactic	70.1	69.9	74.9
IT	Semantic	52.3	54.7	52.3
	Syntactic	51.5	51.8	62.7

https://www.youtube.com/watch?v=CHcExDsDeHU

Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *arXiv preprint arXiv:1607.04606* (2016).

FastText

- Qualitative Results

query	tiling	tech-rich	english-born	micromanaging	eateries	dendritic
ours	tile flooring	tech-dominated tech-heavy	british-born polish-born	micromanage micromanaged	restaurants eaterie	dendrite dendrites
skipgram	bookcases built-ins	technology-heavy .ixic	most-capped ex-scotland	defang internalise	restaurants delis	epithelial p53

https://www.youtube.com/watch?v=CHcExDsDeHU

Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *arXiv preprint arXiv:1607.04606* (2016).