Neural Word Embedding as Implicit Matrix Factorization

(Levy & Goldberg, 2014 NIPS)

Hyunjoong Kim

soy.lovit@gmail.com

github.com/lovit

- Word embedding 의 의미에 대한 해석 관점을 이야기합니다.
  - Negative sampling 을 이용하는 Skipgram 은 word – context 행렬에 Shifted Positive Point Mutual Information (SPPMI) 를 적용한 것과 같고
  - “word – context + SPPMI” 에 SVD 와 같은 차원축소 방법을 적용하면 distributed word representation 을 얻을 수 있습니다.

* Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Advances in neural information processing systems (pp. 2177-2185).

Point Mutual Information (PMI)

- 확률 이론에서는 두 확률이 서로 독립인지 판단하는 방법을 제공합니다.
  - 전체 공간에서의 $p(y)$ 와 x 조건에서의 $p(y|x)$ 가 같으면 x, y 는 독립입니다.

$$\frac{p(x,y)}{p(x) \times p(y)} = \frac{p(y|x)}{p(y)} = 1$$

Point Mutual Information (PMI)

- 확률 이론에서는 두 확률이 서로 독립인지 판단하는 방법을 제공합니다.
  - 두 확률이 독립이면 다음 조건이 성립합니다.

$$\frac{p(x,y)}{p(x)\times p(y)} = 1 \text{ , } \frac{p(\text{안경 } o,\text{저녁 } o)}{p(\text{안경 } o)\times p(\text{저녁 } o)} = \frac{\frac{1}{12}}{\frac{3}{12}\times\frac{4}{12}} = 1$$

	저녁을 먹었다	저녁을 먹지 않았다	Prob.
안경을 썼다	100	200	3 / 12
안경을 쓰지 않았다	300	600	9 / 12
Prob	4 / 12	8 / 12	

Point Mutual Information (PMI)

- 서로 양의 상관성이 있으면 $\frac{p(x,y)}{p(x)\times p(y)}$ 이 1보다 큽니다.

$$\frac{p(x,y)}{p(x)\times p(y)} = 1 \text{ , } \frac{p(\text{안경 } o,\text{저녁 } o)}{p(\text{안경 } o)\times p(\text{저녁 } o)} = \frac{\frac{2}{12}}{\frac{5}{12}\times\frac{3}{12}} = 1.2$$

	저녁을 먹었다	저녁을 먹지 않았다	Prob.
안경을 썼다	200	100	3 / 12
안경을 쓰지 않았다	300	600	9 / 12
Prob	5 / 12	7 / 12	

Point Mutual Information (PMI)

- PMI 는 두 경우의 상관성을 표현하는 index 입니다.

$$PMI(x, y) = \log \frac{p(x, y)}{p(x) \times p(y)}$$

  - 양의 상관관계라면 0 보다 큰 값을 반대라면 0 보다 작은 값을 지닙니다.
  - 값의 방향성에 해석력이 있습니다.

Positive PMI (PPMI)

- 자연어처리에서의 semantic 에서는 음의 상관관계에 큰 의미가 없습니다.
  - 양의 상관관계의 패턴을 강조하기 위해 0 보다 작은 값을 0 으로 변환합니다.

$$PPMI(x, y) = \max\big(0, PMI(x, y)\big)$$

Shifted PPMI

- Shifted PPMI 는 k 보다 큰 PMI 만 값을 보존하며,
  이보다 작은 값은 0 으로 변환합니다.

$$SPPMI_k(x, y) = \max(0, PMI(x, y) - \log(k))$$

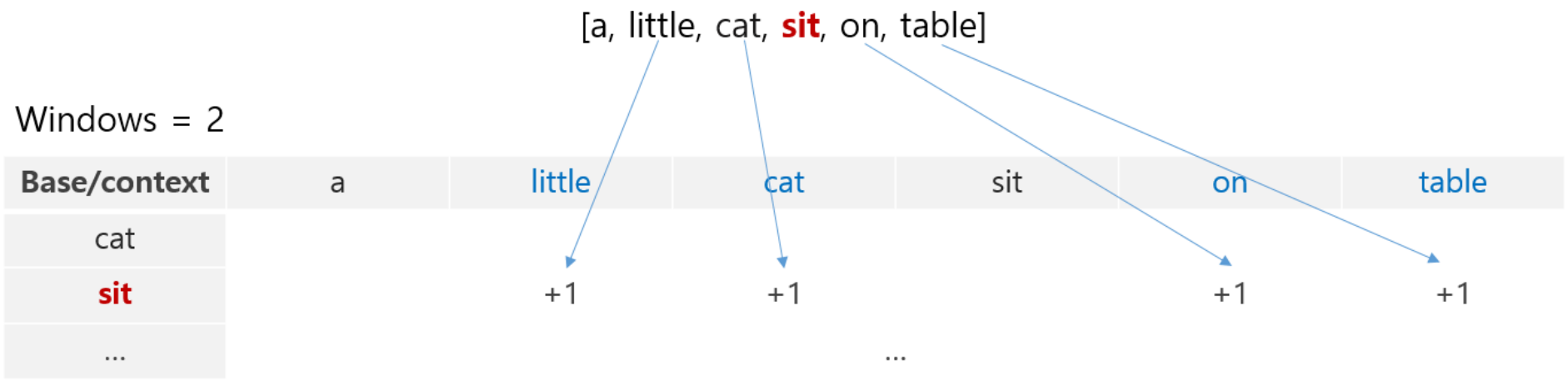
- 논문에서는 negative sampling + Skipgram 의 loss function 을 정리하여 word vector 와 context vector 의 내적이 SPPMI 와 같음을 증명합니다.

$$l = \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c) \cdot \left(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot E_{C_N P_D} [\log \sigma(-\vec{w} \cdot \vec{c_N})] \right)$$

$$\vec{w} \cdot \vec{c} = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \cdot \frac{1}{k} \right) = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log(k)$$

Defining contexts

- (word – context) pair 의 context 는 word 와 앞/뒤로 windows 안에 함께 등장한 단어입니다.



- 논문의 주장을 재현하기 위하여 2016-10-20 뉴스에 대하여 논문의 방법을 적용하였습니다.

Similar words using context vector

- Context vector 만으로도 비슷한 문맥의 단어가 표현됩니다.

query = 박근혜	query = 이화여대	query = 아이오아이	query = 아프리카
('박', 0.847)	('이대', 0.729)	('트와이스', 0.581)	('국가', 0.778)
('두테르테', 0.809)	('사퇴했지만', 0.548)	('블랙핑크', 0.546)	('회사', 0.775)
('노', 0.795)	('사퇴한', 0.520)	('에이핑크', 0.542)	('과거', 0.773)
('아키노', 0.776)	('교수', 0.510)	('아프리카', 0.541)	('일부', 0.753)
('오바마', 0.744)	('사임했습니다', 0.509)	('주제', 0.540)	('변화', 0.749)
('노무현', 0.740)	('총장', 0.506)	('현재', 0.539)	('문제', 0.733)
('박정희', 0.731)	('학교', 0.504)	('이야기', 0.534)	('관계', 0.733)
('백악관에', 0.715)	('최', 0.499)	('경우', 0.533)	('경우', 0.726)
('방북', 0.712)	('정씨', 0.492)	('과거', 0.529)	('이야기', 0.723)
('올랑드', 0.704)	('정유라씨', 0.487)	('태도', 0.527)	('도시', 0.721)

Similar words using context vector + PMI

- 품질이 좋아집니다. '아이오아이'의 유사어에서 '아프리카'가 사라집니다.
  '아프리카'의 유사어도 더 납득이 됩니다.

query = 박근혜	query = 이화여대	query = 아이오아이	query = 아프리카
('대통령', 0.849)	('이대', 0.905)	('신용재', 0.809)	('남미', 0.525)
('박', 0.795)	('최경희', 0.845)	('오블리스', 0.807)	('유럽', 0.510)
('정권', 0.722)	('이화여자대학교', 0.837)	('불독의', 0.773)	('중남미', 0.503)
('국정', 0.713)	('특혜', 0.818)	('백퍼센트', 0.765)	('중동', 0.498)
('비선', 0.676)	('사퇴했지만', 0.811)	('몬스', 0.762)	('호주', 0.496)
('정권의', 0.666)	('총장은', 0.809)	('너무너무', 0.759)	('아시아', 0.482)
('비선실세', 0.661)	('정씨', 0.798)	('갓세븐', 0.755)	('대만', 0.475)
('실세', 0.645)	('총장이', 0.795)	('타이틀곡', 0.752)	('오세아니아', 0.470)
('수석비서관회의에서', 0.640)	('입학', 0.790)	('엠카운트다운', 0.748)	('케냐', 0.465)
('최순실', 0.633)	('특혜입학', 0.783)	('불독은', 0.744)	('해외', 0.463)

Similar words using context vector + PMI + SVD

- SVD 의 차원을 300, 30, 10 으로 줄여가며 유사어를 찾아봅니다.

query=박근혜, SVD d=300	query=박근혜, SVD d=30	query=박근혜, SVD d=10
('대통령', 0.849)	('대통령', 0.934)	('박', 0.980)
('박', 0.795)	('국정', 0.911)	('민주당', 0.978)
('정권', 0.722)	('박', 0.911)	('새누리당', 0.978)
('국정', 0.713)	('정권', 0.909)	('대통령', 0.976)
('비선', 0.676)	('청와대', 0.897)	('문재인', 0.974)
('정권의', 0.666)	('게이트', 0.884)	('대표가', 0.972)
('비선실세', 0.661)	('김대중', 0.871)	('우', 0.967)
('실세', 0.645)	('정권의', 0.867)	('원장은', 0.966)
('수석비서관회의에서', 0.640)	('파문', 0.845)	('국민의당', 0.965)
('최순실', 0.633)	('수석', 0.844)	('더민주', 0.964)

Similar words using context vector + PMI + SVD

- SVD 의 차원을 300, 30, 10 으로 줄여가며 유사어를 찾아봅니다.

query=이화여대, SVD d=300	query=이화여대, SVD d=30	query=이화여대, SVD d=10
('이대', 0.905)	('이대', 0.972)	('원장', 0.987)
('최경희', 0.845)	('최경희', 0.950)	('이대', 0.980)
('이화여자대학교', 0.837)	('총장은', 0.946)	('총장', 0.978)
('특혜', 0.818)	('입학', 0.932)	('정', 0.977)
('사퇴했지만', 0.811)	('학사', 0.930)	('최', 0.971)
('총장은', 0.809)	('총장이', 0.929)	('변호사', 0.970)
('정씨', 0.798)	('이화여자대학교', 0.928)	('총장은', 0.959)
('총장이', 0.795)	('정씨', 0.921)	('최경희', 0.957)
('입학', 0.790)	('사퇴했지만', 0.917)	('위원장', 0.956)
('특혜입학', 0.783)	('특혜', 0.916)	('현직', 0.951)

Similar words using context vector + PMI + SVD

- 차원이 작을수록 (minor components 를 이용하지 않을수록) 유사어 좋지 않습니다.

query=아이오아이, SVD d=300	query=아이오아이, SVD d=30	query=아이오아이, SVD d=10
('신용재', 0.809)	('몬스', 0.958)	('라디오스타', 0.989)
('오블리스', 0.807)	('샤이니', 0.957)	('불독', 0.989)
('불독의', 0.773)	('불독은', 0.949)	('불독은', 0.987)
('백퍼센트', 0.765)	('멤버', 0.941)	('예능프로그램', 0.987)
('몬스', 0.762)	('불독의', 0.939)	('몬스', 0.987)
('너무너무', 0.759)	('불독', 0.936)	('사랑하기', 0.986)
('갓세븐', 0.755)	('타엑스', 0.929)	('한끼줍쇼', 0.986)
('타이틀곡', 0.752)	('엑소', 0.929)	('구르미', 0.985)
('엠카운트다운', 0.748)	('타이틀곡', 0.928)	('주연을', 0.984)
('불독은', 0.744)	('너무너무', 0.928)	('달빛', 0.984)

Similar words using context vector + PMI + SVD

- 차원이 작을수록 (minor components 를 이용하지 않을수록) 유사어 좋지 않습니다.

query=아프리카, SVD d=300	query=아프리카, SVD d=30	query=아프리카, SVD d=10
('남미', 0.525)	('남미', 0.897)	('코', 0.973)
('유럽', 0.510)	('호주', 0.882)	('에선', 0.971)
('중남미', 0.503)	('유일한', 0.881)	('엘리', 0.971)
('중동', 0.498)	('자국', 0.880)	('전쟁', 0.969)
('호주', 0.496)	('에선', 0.880)	('비롯', 0.966)
('아시아', 0.482)	('이스라엘', 0.878)	('에서도', 0.963)
('대만', 0.475)	('대만', 0.877)	('인도', 0.963)
('오세아니아', 0.470)	('노리는', 0.874)	('박사는', 0.961)
('케냐', 0.465)	('칠레', 0.874)	('계에', 0.961)
('해외', 0.463)	('인도', 0.874)	('성공을', 0.960)

Comparison

Word2Vec	PMI + SVD_300	norm(Context) + SVD_300	Context + SVD_300	Context + PMI
(‘노무현’, 0.780)	(‘대통령’, 0.849)	(‘위도도’, 0.885)	(‘위도도’, 0.896)	(‘대통령’, 0.334)
(‘테메르’, 0.774)	(‘박’, 0.795)	(‘노’, 0.883)	(‘두테르테’, 0.885)	(‘박’, 0.185)
(‘연설문’, 0.765)	(‘정권’, 0.722)	(‘두테르테’, 0.872)	(‘노’, 0.884)	(‘정권’, 0.156)
(‘오바마’, 0.762)	(‘국정’, 0.713)	(‘오바마’, 0.869)	(‘아키노’, 0.879)	(‘최순실’, 0.153)
(‘박’, 0.753)	(‘비선’, 0.676)	(‘올랑드’, 0.863)	(‘오바마’, 0.875)	(‘청와대’, 0.146)
(‘두테르테’, 0.709)	(‘정권의’, 0.666)	(‘박’, 0.857)	(‘올랑드’, 0.873)	(‘정부의’, 0.128)
(‘연설문을’, 0.708)	(‘비선실세’, 0.661)	(‘블라디미르’, 0.848)	(‘국가안보실’, 0.865)	(‘비선’, 0.127)
(‘위도도’, 0.704)	(‘실세’, 0.645)	(‘국가안보실’, 0.845)	(‘박’, 0.864)	(‘정권의’, 0.126)
(‘노’, 0.696)	(‘수석비서관회의에서’, 0.640)	(‘방북’, 0.841)	(‘테메르’, 0.857)	(‘정부’, 0.118)
(‘국가안보실과’, 0.682)	(‘최순실’, 0.633)	(‘아키노’, 0.840)	(‘블라디미르’, 0.855)	(‘의혹’, 0.113)

Comparison

Word2Vec	PMI + SVD_300	norm(Context) + SVD	Context + SVD	Context + PMI
('에이핑크', 0.833)	('신용재', 0.809)	('트와이스', 0.817)	('트와이스', 0.863)	('신용재', 0.236)
('샤이니', 0.828)	('오블리스', 0.807)	('에이핑크', 0.790)	('아프리카발톱개구리', 0.834)	('너무너무', 0.232)
('타이틀곡', 0.801)	('불독의', 0.773)	('아프리카발톱개구리', 0.785)	('칼라', 0.830)	('엠카운트다운', 0.226)
('빅스', 0.793)	('백퍼센트', 0.765)	('언니', 0.784)	('언니', 0.825)	('갓세븐', 0.184)
('불독의', 0.788)	('몬스', 0.762)	('주니어', 0.784)	('아프리카', 0.825)	('걸그룹', 0.183)
('엑소', 0.779)	('너무너무', 0.759)	('왕자', 0.776)	('썬코어', 0.821)	('컴백', 0.181)
('트와이스', 0.775)	('갓세븐', 0.755)	('아프리카', 0.771)	('사우어크라우트', 0.821)	('완전체로', 0.178)
('몬스', 0.769)	('타이틀곡', 0.752)	('비버', 0.768)	('에이핑크', 0.818)	('엠카', 0.172)
('다이아', 0.767)	('엠카운트다운', 0.748)	('무대', 0.762)	('스트레인지', 0.818)	('오블리스', 0.169)
('씨엔블루', 0.756)	('불독은', 0.744)	('창원시', 0.761)	('톰', 0.818)	('다비치', 0.167)

Discussion

- Context vector 만으로도 설겁게나마 유사어 검색이 가능합니다.
  - 주위에 등장하는 단어의 분포가 비슷한 두 단어의 의미는 비슷합니다.
  - 하지만 어느 단어에나 함께 등장하는 context words 의 영향력이 큽니다.
- Point Mutual Information 은 특별한 문맥 없이 자주 등장하는 단어의 영향력을 줄여주는 역할을 합니다.
  - 일종의 term weighting 입니다.

Discussion

- Singular Value Decomposition (SVD) 는 sparse context vector 에서 서로 상관성이 높은 context words 를 하나의 축으로 묶음으로써 distributed representation 을 학습합니다.
- 같은 문맥이지만 서로 다른 단어였던 context words 가 하나의 축으로 묶이기 때문에 SVD 를 적용하면 전체적인 유사도가 증가합니다.

Conclusion

- Representation 을 위해서는 descriptor 가 무엇인지를 잘 정의하는 것이 중요합니다.
- (word, context) pairs 처럼 distributed representation 으로 표현할 대상의 descriptor 를 잘 설정하면 PMI + SVD 와 같은 방법으로도 representation 을 학습할 수 있습니다.

References

- https://lovit.github.io/nlp/2018/04/22/context_vector_for_word_similarity/