GloVe(Global Vectors for Word Representation)

박주찬

DICE Lab
School of Computer Science and Engineering
KOREATECH
green261535@gmail.com

Introduction

GloVe(Global Vectors for Word Representation)는 count based와 direct predict를 모두 사용하는 방법론으로 2014년에 미국 스탠포드대학에서 개발한 단어 임베딩 방법론이다. Count based의 LSA(Latent Semantic Analysis)와 direct predict의 Word2Vec에 단점을 지적하며 이를 보완한다는 목적으로 나왔다. 실제로도 GloVe는 Word2Vec만큼이나 뛰어난 성능을 보여준다. 단정적으로 Word2Vec와 GloVe 중에서 어떤 것이 더 뛰어나다고 말할 수는 없고, 이 두 가지 전부를 사용해보고 성능이 더 좋은 것을 사용하는 것이 바람직하다고 한다.

Count based와 direct prediction

Count based에서 LSA는 문서에서의 각 단어의 빈도수를 카운트 한 행렬이다. Count based는 전체적인 통계 정보를 입력으로 받아 차원을 축소(Truncated SVD)하여 잠재된 의미를 끌어내는 방법론이었다. 반면 direct prediction에서 Word2Vec는 실제값과 예측값에 대한 오차를 loss function을 통해 줄여 나가며 학습하는 direct predict 방법론이었다. 서로 다른 방법을 사용하는 이 두 방법론은 각각 장, 단점이 있다. LSA는 count based로 corpus의 전체적인 통계 정보를 고려하지만, 왕:남자 = 여왕:? (정답은 여자)와 같은 단어 의미의 유추 작업(Analogy task)에는 성능이 떨어진다. Word2Vec는 direct prediction로 단어 간 유추 작업에는 LSA보다 뛰어나지만, 임베딩 벡터가 window size 내에서만 주변 단어를 고려하기 때문에 corpus의 전체적인 통계 정보를 반영하지 못합니다. GloVe는 이러한 기존 방법론들의 각각의 한계를 지적하며, LSA의

메커니즘이었던 count based 방법과 Word2Vec의 메커니즘이었던 direct prediction 방법론 두 가지를 모두 사용한다.

Window based Co-occurrence Matrix

Example corpus:

- I like deep learning.
- I like NLP.
- I enjoy flying.

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

Window size = 1일 때, 위의 텍스트를 가지고 구성한 Co-occurrence Matrix이다. 위 행렬은 행렬을 전치 해도 동일한 행렬이 된다는 특징이 있다.

Co-occurrence Probability

Co-occurrence probability P(k|i)는 Co-occurrence Matrix로부터 특정 단어 i의 전체 등장 횟수를 카운트하고, 특정 단어 i가 등장 했을 때 어떤 단어 k가 등장한 횟수를 카운트하여 계산한 조건부 확률이다. P(k|i)에서 i를 center word, k를 context word라고 했을 때, 위 Co-occurrence Matrix의 표에서 center word i의 행의 모든 값을 더한 값을 분모로 하고, i행 k열의 값을 분자로 한 값이라고 볼 수 있다.

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

위의 표를 통해 알 수 있는 사실은 solid가 등장했을 때 ice가 등장할 확률 0.00019은 solid가 등장했을 때 steam이 등장할 확률인 0.000022보다 약 8.9배 크다는 것이다. 쉽게 생각해보면 solid는 '단단한'이라는 의미를 가졌으니까 '증기'라는 의미를 가지는 steam보다는 당연히 '얼음'이라는 의미를 가지는 ice라는 단어와 더 자주 등장할 것이다.

수식적으로 다시 정리하여 언급하면 k가 solid일 때, P(solid l ice) / P(solid l steam)를 계산한 값은 8.9가 나온다. 이 값은 1보다는 매우 큰 값이다. 왜냐면 P(solid | ice)의 값은 크고, P(solid | steam)의 값은 작기 때문이다. 그런데 k를 solid가 아니라 gas로 바꾸면 값은 완전히 달라진다. gas는 ice보다는 steam과 더 자주 등장하므로, P(gas l ice) / P(gas l steam)를 계산한 값은 1보다 훨씬 작은 값인 0.085가 나오게 된다. 반면, k가 water인 경우에는 solid와 steam 두 단어 모두와 동시 등장하는 경우가 많으므로 1에 가까운 값이 나오고, k가 fashion인 경우에는 solid와 steam 두 단어 모두와 동시 등장하는 경우가 적으므로 1에 가까운 값이 나온다.

Loss Function

GloVe의 아이디어를 한 줄로 요약하면 '**임베딩 된 중심 단어와 주변 단어 벡터의 내적이 전체 코퍼스에서의 동시 등장 확률이 되도록 만드는 것**'이다. 즉, 아래의 식을 만족하도록 임베딩 벡터를 만드는 것이 목표이다. w_i는 중심 단어 i의 임베딩 벡터, $\tilde{w}_k$는 주변 단어 k의 임베딩 벡터이다.

$$\text{dot product}(w_i, \tilde{w}_k) \approx P(k|i) = P_{ik}$$

위 식에서 좌변 dot product는 음수가 될 수도 있고 양수가 될 수 있지만 우변은 확률 값이므로 항상 양수이다. 그래서 우변에 log를 취해 값의 범위를 좌변과 맞춰준다. 그러므로 아래와 같은 관계를 가지도록 임베딩 벡터를 설계한다.

$$\text{dot product}(\mathrm{w_i},\overline{\mathrm{w}}_\mathrm{k}) \approx \log\big(\mathrm{P(k|i)}\big) = \log(\mathrm{P_{ik}})$$

GloVe의 연구진들은 벡터 $w_i, w_j, \overline{w}_k$를 가지고 어떤 함수 F를 수행하면, $\frac{P_{ik}}{P_{jk}}$ 가 나온다는 아래와 같은 초기 식으로부터 전개를 시작한다.

$$F\big(w_i, w_j, \overline{w}_k\big) = \frac{P_{ik}}{P_{jk}}$$

함수 F는 두 단어 사이의 Co-occurrence probability 크기 관계 비(ratio) 정보를 벡터 공간에 인코딩하는 것이 목적이다. 이를 위해 GloVe 연구진들은 w_i 와 w_j 라는 두 벡터의 차이를 함수 F의 입력으로 사용하는 것을 제안한다.

$$F\big((w_i - w_j), \overline{w}_k\big) = \frac{P_{ik}}{P_{jk}}$$

위 식에서 우변은 확률이므로 스칼라 값이다고, 좌변은 벡터 값이다. 이를 성립하기 해주기 위해서 함수 F는 두 입력에 내적을 수행한다.

$$F\big((w_i - w_j)^T \overline{w}_k\big) = \frac{P_{ik}}{P_{jk}}$$

GloVe 연구팀에 따르면 여기까지 도출된 F는 다음 세가지 조건을 만족해야 한다. 우선 w_i 와 $\overline{w}_k$를 서로 바꾸어도 식이 같은 값을 반환해야 한다. 왜냐하면 context word $\overline{w}_k$는 얼마든지 center word w_i나 w_j가 될 수 있기 때문이다. 또한 corpus 전체에서 구한 co-occurrence matrix X는 대칭 행렬이므로 두번째 성질을 포함해야 한다. 마지막으로 homomorphism(준동형 사상)조건을 만족해야 하므로 세번째 성질을 만족해야한다.

$$w_i \leftrightarrow \overline{w}_k$$

$$\mathrm{X} \leftrightarrow X^T$$

$$\mathrm{F(X\text{-}Y)} = \frac{\mathrm{F(X)}}{\mathrm{F(Y)}}$$

위 세번째 성질을 GloVe에 적용시키면 함수 F의 우변은 다음과 같이 바뀐다.

$$F\left(\left(w_i - w_j\right)^T \bar{w}_k\right) = \frac{P_{ik}}{P_{jk}}$$

$$F\left({w_i}^T \bar{w}_k - {w_j}^T \bar{w}_k\right) = \frac{F({w_i}^T \bar{w}_k)}{\mathrm{F}({w_j}^T \bar{w}_k)}$$

이러한 조건을 만족하는 함수 F는 지수 함수이기 때문에 함수 F를 exp로 치환하고 식을 정리하면 아래와 같다.

$$\begin{aligned} exp({w_i}^T \bar{w}_k) &= P_{ik} \\ &= \frac{X_{ik}}{X_i} \end{aligned}$$

$\mathrm{P}(k|i) = P_{ik} = \frac{X_{ik}}{X_i}$ 중심단어 i가 등장했을 때 윈도우 내 주변 단어 k가 등장할 확률이다. 위 식에 양변에 log를 취하면 아래와 같다.

$$\begin{aligned} {w_i}^T \bar{w}_k &= \log(P_{ik}) \\ &= \log\left(\frac{X_{ik}}{X_i}\right) \\ &= \log(X_{ik}) - \log(X_i) \end{aligned}$$

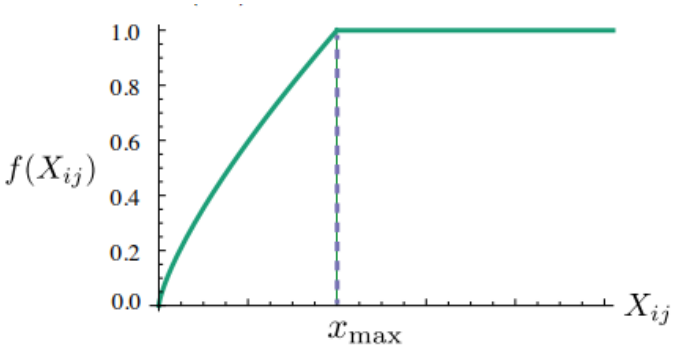
그런데 여기서 앞에 F 함수는 세가지 성질을 만족해야 한다고 했듯이, w_i 와 $\bar{w}_k$ 는 두 값의 위치를 서로 바꾸더라도 식이 성립해야 한다. 그러기 위해서는 $\log(P_{ik}) = \log(P_{ki})$ 를 만족해야한다. 결국 $\log(X_{ik}) - \log(X_i) = \log(X_{ki}) - \log(X_k)$ 를 만족해야한다. $\log(X_{ik}) = \log(X_{ki})$ 이지만, $\log(X_i)$ 와 $\log(X_k)$가 다르기 때문에 $\log(X_{ik}) - \log(X_i)$와 $\log(X_{ki}) - \log(X_k)$는 서로 다르다. 이러한 이유들로, GloVe 연구팀은 $\log(X_i)$를 아래와 같이w_i에 대한 bias term b_i라는 상수항으로 대체한다. 같은 이유로 $\bar{w}_k$에 대한 bias term $\widetilde{b_k}$를 추가하면 아래와 같은 식이 나오는 것을 볼 수 있다.

$$\begin{aligned} {w_i}^T \bar{w}_k &= \log(X_{ik}) - b_i - \widetilde{b_k} \\ {w_i}^T \bar{w}_k + b_i + \widetilde{b_k} &= \log(X_{ik}) \end{aligned}$$

이 식이 손실 함수의 핵심이 되는 식이며, 우변의 값과의 차이를 최소화 하는 방향으로 좌변의 4개의 항은 학습을 통해 값이 바뀌는 변수들이 된다. Loss function을 아래와 같이 정의 할 수 있다.

$$\text{Loss Function} = \sum_{i,j=1}^{V} \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log(X_{ij})\right)^2$$

여기서 V는 단어 집합의 크기를 의미한다. 그런데 아직 최적의 손실 함수라기에는 부족하다. GloVe연구진은 $\log(X_{ik})$ 에서 X_{ik} 값이 0이 되면 $\log(X_{ik})$값은 음의 무한대가될 수 있음을 지적한다. 대안 중 하나는 $\log(X_{ik})$항을 $\log(1 + X_{ik})$로 변경하는 것이다. 하지만 이렇게 해도 여전히 해결되지 않는 문제가 있다. 바로 co-occurrence matrix X는 희소 행렬일 가능성이 높다는 점이다. X에는 많은 값이 0이거나, 동시 등장 빈도가 적어서 많은 값이 작은 수치를 가지는 경우가 많다. GloVe 연구진들은 co-occurrence matrix에서 동시 등장 빈도의 값 X_{ik}가 굉장히 낮은 경우에는 정보에 거의 도움이 되지 않는다고 판단한다. 그래서 이에 대한 가중치를 주는 고민을 하게 되어 X_{ik}의 값에 영향을 받는 가중치 함수 $f(X_{ik})$ 를 손실 함수에 도입하게 된다. GloVe에 도입되는$f(X_{ik})$의 그래프를 보면 아래와 같다.



X_{ik}의 값이 작으면 상대적으로 함수의 값은 작도록 하고, 값이 크면 함수의 값은 상대적으로 크도록 한다. 하지만 X_{ik}가 지나치게 높다고 해서 지나친 가중치를 주지 않기 위해서 함수의 최대값(최대값=1)이 정해져 있다. 예를 들어서 자주 나오는 ‘it is’와 같은 불용어의 등장 빈도수가 높다고 해서 지나친

가중을 받아서는 안되기 때문이다. $f(X_{ik})$이 함수의 값을 손실 함수에 곱해주면 가중치의 역할을 할 수 있다. $f(X_{ik})$의 식은 아래와 같이 정의 한다.

$$f(x) = \min(1, \left(\frac{\mathrm{x}}{x_{max}}\right)^{\frac{3}{4}})$$

최종적으로 다음과 같은 일반화 된 손실 함수를 얻을 수 있다.

$$Loss\ function = \sum_{i,j=1}^{V} f(X_{ij})(w_i{}^T \overline{w}_j + b_i + \widetilde{b_j} - \log(X_{ij}))^2$$

Conclusion

GloVe는 임베딩 된 center word와 context word 벡터의 내적이 전체 corpus에서의 co-occurrence probability가 되도록 만든 것 즉, count based의 장점과 direct prediction의 장점을 가져와서 기존 방법론의 단점을 보완한 것이 성능향상에 도움을 주었다고 생각한다.

References

1. https://wikidocs.net/book/2155
2. https://ratsgo.github.io/