GloVe:
Global Vectors forWord Representation

ABSTRACT

- 단어의 벡터 공간 표현을 학습하는 최근의 방법은 벡터 산술을 사용하여 세분화 된 의미 및 구문 규칙성을 포착하는 데 성공했지만 이러한 규칙성의 근원은 여전히 불투명합니다.

- 우리는 그러한 어휘들이 단어 벡터에 나타나기 위해 필요한 모델 속성을 분석하고 명시한다.

- 그 결과는 문헌에서 두 가지 주요 모델 군의 이점,
  즉 global matrix factorization 및 local context window methods 방법을 결합한 **새로운 logbilinear regression model입니다.**

- 우리의 모델은 전체 희소 행렬 또는 대형 코퍼스의 개별 컨텍스트 window보다는 **word-word co-occurrence matrix에서 0이 아닌 요소에 대해서만 학습함으로써** 통계 정보를 효율적으로 활용합니다.

- 이 모델은 최근 단어 유추 작업에서 75 %의 성능으로 입증 된 의미있는 하위 구조가있는 벡터 공간을 생성합니다. 또한 유사성 작업 및 명명 된 엔티티 인식에 대한 관련 모델을 능가합니다.

INTRODUCTION

- 언어의 semantic 벡터 공간 모델은 각 단어를 실수 벡터로 나타냅니다
- 이러한 벡터는 정보 검색 (Manning et al., 2008), 문서 분류 (Sebastiani, 2002), 질의 응답 (Tellex et al., 2003), 지명 된 개체 인식 (Turian et al., 2010) 및 파싱 (Socher et al., 2013)등에 사용됩니다.
- 대부분의 단어 벡터 방법은 이러한 단어 표현 집합의 본질적 품질을 평가하기위한 기본 방법으로 단어 벡터 pair 사이의 거리 또는 각도에 의존합니다.
- 최근, Mikolov et al. (2013c)는 단어 벡터 간의 스칼라 거리가 아니라 다양한 차이점을 조사하여 단어 벡터 공간의 미세 구조를 조사하는 단어 유추에 기반한 새로운 평가 체계를 도입했습니다.
  예를 들어, "왕은 남자가 여자로 여왕이된다."이라는 비유는 벡터 방정식 king - queen = man - woman으로 벡터 공간에 인코딩되어야 합니다. 이 평가 방법은 의미의 차원을 생성하는 모델을 선호하므로 분산 표현의 다중 클러스터링 아이디어를 캡처합니다 (Bengio, 2009).
- The two main model
  1) global matrix factorization methods, such as latent semantic analysis (LSA) (Deerwester et al., 1990)
  2) local context window methods, such as the skip-gram model of Mikolov et al. (2013c).
- 현재 두 가정 모두 심각한 결점을 가지고 있습니다
- LSA는 통계 정보를 효율적으로 활용하며, 유사성이 낮은 단어 공간 작업에 상대적으로 부적절하며 하위 최적의 벡터 공간 구조를 나타냅니다
  - LSA는 개념적으로 co-occurrence 정보를 이용한다. co-occurrence 정보를 이용한다는 것은 단어의 '형태(morphology)가 아닌 의미(semantic)'를 이용한다는 뜻이다. 예를 들어 '배'라는 단어는 같은 문장에 co-occur 하는 동사가 '타다' 인지 '먹다' 인지에 따라 의미가 갈라지게 된다. 또한, '식당', '맛있게', '배부르게' 라는 단어와 같은 문장에 co-occur하는 처음 보는 단어는 아마 '음식'일 것이다. LSA는 이론적으로는 선형대수학의 SVD(Singular Value Decomposition)을 이용한다. SVD를 계산하는 방법에 대해서는 얘기하지 않을 것이다.
- skip-gram과 같은 메소드는 유추 작업에서 더 잘할 수 있지만 전역 동시 발생 횟수 대신 별도의 로컬 컨텍스트 창에서 트레이닝하므로 코퍼스 통계를 제대로 활용하지 못합니다.

INTRODUCTION

- 이 연구에서 우리는 의미의 선형 방향을 생성하는 데 필요한 모델 특성을 분석하고 이를 위해 global log-bilinear regression models이 적절하다고 주장한다.

- 우리는 global word-word co-occurrence에 대해 훈련하는 specific weighted least squares model(특정 가중치 최소 제곱) 모델을 제안하여 통계를 효율적으로 사용합니다.

- 이 모형은 유사 유물이라는 단어에 대한 75 %의 정확성을 자랑하는 최첨단 성능으로 입증 된 의미있는 하부 구조를 가진 단어 벡터 공간을 생성합니다.

- 우리는 또한 우리의 방법이 몇 가지 단어 유사성 작업에 대한 현재의 다른 방법과 공통의 명명 된 엔티티 인식(NER) 벤치 마크에서 우위에 있음을 입증합니다.

RELATED WORK

- Matrix Factorization Methods. (행렬 인수 분해 방법.)
  - 저 차원 단어 표현을 생성하기위한 행렬 인수 분해 방법은 LSA만큼 거슬러 올라간 뿌리를 가지고 있습니다
  - 이러한 방법은 낮은 순위의 근사를 사용하여 corpus에 대한 통계 정보를 캡처하는 큰 행렬을 분해합니다.
  - LSA에서, 행렬은 단어 또는 용어에 대응하는 "term-document"유형의 행렬이며, 열은 코퍼스의 다른 문서에 해당합니다.

THE GLOVE MODEL

- corpus에서 단어 발생의 통계는 단어 표현을 학습하기 위한 모든 unsupervised methods에 사용할 수있는 주요 정보원이며, 현재 그러한 방법이 많이 있지만 이러한 통계로 인해 의미가 생성되는 방법에 대한 의문은 여전히 남아 있습니다.

- Let the matrix of **word-word co-occurrence counts be denoted by X**,
  whose entries Xi j tabulate the number of times word j occurs in the context of word i.

- First we establish some notation
  Finally, let **Pi j = P(j ji) = Xi j=Xi** be the probability that word j appear in the

- 우리는 i = ice 및 j = steam을 취할 수있는 열역학 단계의 개념에 관심이 있다고 가정 해보십시오. 이 단어의 관계는 여러 프로브 단어 k와의 동시 발생 확률의 비율을 조사하여 조사 할 수 있습니다. 얼음과 관련이 있지만 증기가 아닌 단어 k에 대해 k = solid라고하면 Pik / Pjk의 비율이 클 것으로 예상됩니다. 유사하게, 증기와 관련이 있지만 얼음이 아닌 단어 k에 대해, k = 가스라고하면, 그 비율은 작아야한다.

- 비율 Pik / Pjk가 세 단어 i, j 및 k에 의존한다는 점을 주목하면 가장 일반적인 모델은 다음과 같은 형식을 취합니다.

- $$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}, \qquad (1)$$

- 이 방정식에서, 오른쪽은 corpus에서 추출되고 F는 몇 가지 미정의 매개 변수에 따라 달라질 수 있습니다.
  F의 확률은 다양하지만, 몇 가지 사항을 강요함으로써 우리는 독특한 선택을 할 수 있습니다.

- F가 단어 벡터 공간에서 Pik / Pjk의 비율로 나타나는 정보를 인코딩 하도록 합니다.
  벡터 공간은 본질적으로 선형 구조이므로이 작업을 수행하는 가장 자연스러운 방법은 벡터 차이입니다.
  이 목적을 가지고, 우리는 두 개의 목표 단어의 차이에만 의존하는 함수 F로 고려를 제한 할 수있다.

THE GLOVE MODEL

- 다음으로 우리는 Eqn. (2)는 벡터이고 오른쪽은 스칼라이다. F는 예를 들어 신경망에 의해 매개 변수화 된 복잡한 함수로 간주 될 수 있지만 그렇게하면 우리가 포착하려고하는 선형 구조를 난독화하게된다.
  이 문제를 피하기 위해 먼저 인수의 내적을 취할 수 있습니다. >> Eqn.(2)

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}. \qquad (2)$$

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}, \qquad (3)$$

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}, \qquad (4)$$

which, by Eqn. (3), is solved by,

$$F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}. \qquad (5)$$

The solution to Eqn. (4) is $F = \exp$, or,

$$w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i). \qquad (6)$$

THE GLOVE MODEL

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik}) . \quad (7)$$

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2 , \quad (8)$$

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} . \quad (9)$$

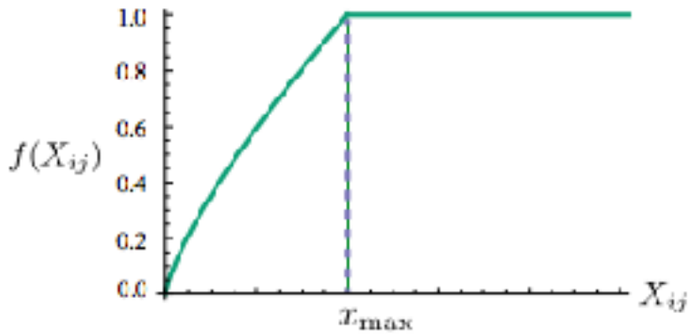

Figure 1: Weighting function f with $\alpha = 3/4$.

$$J(\theta) = \frac{1}{2} \sum_{i,j=1}^{W} f(P_{ij})(u_i^T v_j - \log P_{ij})^2$$

Prediction

How often i&j co-occur

- 동시 발생 행렬의 로그를 인수 분해하는 아이디어는 LSA와 밀접한 관련이 있으며, 결과 모델을 실험의 기준으로 사용합니다.
  이 모델의 가장 큰 단점은 드물게 발생하거나 심지어는 전혀 발생하지 않는 모든 경우에도 동등하게 모든 발생을 비교한다는 것입니다.
  이러한 rare cooccurrences은 소음이 많고 빈번한 정보보다 적은 정보를 담고 있습니다.
  어휘 항목 및 코퍼스에 따라 X의 데이터 중 75 ~ 95 % 만 차지합니다.
- (7)을 **least squares problem** 로 대입하고 비용 함수에 **weighted** 함수 f (Xi j)를 도입하면 모델 (8)이 된다.
  ==> 이것이 문제 해결을 위한 **new weighted least squares regression model**
- (8) V is the size of the vocabulary.
- weighting function should obey the following properties:

  1. f (0) = 0. If f is viewed as a continuous function, it should vanish as x → 0 fast enough that the limx→0 f (x) log2 x is finite.

  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 that frequent co-occurrences are not overweighted.
- 모델의 성능은 cutoff에 약하게 의존,
  모든 실험에서 xmax = 100으로 고정
  α = 3/4가 **α** = 1 인 선형 버전에 비해 완만한 개선을 발견
  3/4 값을 선택하는데 경험적 동기만을 제공하지만, 비슷한 분수의 전력 스케일링이 가장 좋았던 것이 흥미 롭다.

THE GLOVE MODEL - SUMMARY

- The GloVe model is trained on the non-zero entries of a global word-word co-occurrence matrix, which tabulates how frequently words co-occur with one another in a given corpus.
  (Glove model은 얼마나 단어가 자주 서로 발생하는지에 대한
  global word-word co-occurrence matrix의 0이 아닌 요소를 학습한다)

- Subsequent training iterations are much faster because the number of non-zero matrix entries is typically much smaller than the total number of words in the corpus.
  (0이 아닌 행렬 항목의 수가 일반적으로 코퍼스의 총 단어 수보다 훨씬 적기 때문에 다음 학습 반복이 훨씬 빠르다.)

- GloVe is essentially a log-bilinear model with a weighted least-squares objective.

- The training objective of GloVe is to learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence.
  (단어의 벡터의 내적이 단어의 동시 발생 확률의 로그와 같도록 학습하는 것이다.)

- logarithm of a ratio = the difference of logarithms (이것이 같기 때문에)
  this objective associates (the logarithm of) ratios of co-occurrence probabilities with vector differences in the word vector space. (objective는 벡터공간의 벡터차이를 가진 co-occurrence probabilities에 연관 시킨다.)

- these ratios can encode some form of meaning => encoded as vector differences as well
  어떤 의미의 형태를 인코딩 할수 있는 비율은 벡터의 차이를 더 좋게 인코딩 가능하다.

- For this reason, the resulting word vectors perform very well on word analogy tasks, such as those examined in the word2vec package.
  (이것은 word2Vec package에서 검사된 것과 같이, 단어 유추의 결과에 좋은 결과가 있다)

CONCLUSION

- Currently, prediction-based models garner substantial support
- In this work we argue that the two classes of methods are not dramatically different at a fundamental level
  - they both probe the underlying co-occurrence statistics of the corpus
    (두개의 연구가 기존 방법과 극적으로 차이가 있지는 않음)
  - **but the efficiency with which the count-based methods capture global statistics can be advantageous**
    (count-based의 이점)
- We construct a model that utilizes this main benefit of count data while simultaneously capturing the meaningful linear substructures prevalent in recent log-bilinear prediction-based methods like word2vec.
  (Word2vec와 같은 최근 로그 쌍 선형 예측 기반 방법에서 널리 사용되는 의미있는 선형 하부 구조를 동시에 캡처하면서 카운트 데이터의 이러한 주요 이점을 활용하는 모델을 만듦)
- GloVe, is a new global log-bilinear regression model for the unsupervised learning of word representations that outperforms other models on word analogy, word similarity, and named entity recognition tasks.
  (단어 유추, 단어 유사성, 명명된 엔티티 인식 작업에서 다른 모델보다 우수한 단어 표현의 Unsupervised Learning을 위한 new global log-bilinear regression model)

HOW IS GLOVE DIFFERENT FROM WORD2VEC?

- They differ in that word2vec is a "predictive" model,
  whereas GloVe is a "count-based" model
- 예측 모델은 손실의 예측 능력을 향상시키기 위해 벡터를 학습
  Loss(target word | context words; Vectors
- In word2vec, feed-forward neural network and optimized SGD
- Count-based models learn their vectors by essentially doing dimensionality reduction on the co-occurrence counts matrix.
  매트릭스에서 차원 감소를 수행하여 벡터를 학습

- first construct a large matrix of (words x context) co-occurrence information
- "word" (the rows), you count how frequently
  we see this word in some "context" (the columns)
  (각 word (행)에 대해 큰 corpus의 "context"(열)에서이 단어를 얼마나 자주 볼 수 있는지 계산)
- The number of "contexts" is of course large
  (컨텍스트는 본질적으로 크기가 큰 조합)
- they factorize this matrix to yield a lower-dimensional (word x features) matrix
  (행렬을 인수분해 하여 낮은 차원의 (단어 X 특징) 행렬을 산출)
- each row now yields a vector representation for each word
  (각 행은 각 단어에 대한 벡터 표현을 산출)

- http://nlp.stanford.edu/pubs/glove.pdf
- https://www.quora.com/How-is-GloVe-different-from-word2vec
- https://en.wikipedia.org/wiki/Log-linear_model
- http://clic.cimec.unitn.it/marco/publications/acl2014/baroni-etal-countpredict-acl2014.pdf
- http://sragent.tistory.com/entry/Latent-Semantic-AnalysisLSA