Glove: Global Vectors for Word Representation

https://www.researchgate.net/publication/284576917_Glove_Global_Vectors_for_Word_Representation

0. Introduction

- NLP에서 단어 의미를 벡터로 표현하는 방법은 중요한 연구 주제
- 기존 Word2Vec은 local context(주변 단어 정보)만 활용하여 벡터를 학습
- 본 논문은 전역 통계(Global Co-occurrence Statistics)를 활용해 단어 벡터를 학습하는 새로운 방법 제안
- 목표 : 단어 벡터가 의미 관계를 보다 정확하게 반영하도록 함

1. Overview

- GloVe : 단어 공동 출현(co-occurrence) 행렬 기반의 단어 임베딩 기법
- 특징:
  - 전체 말뭉치에서 단어 간 통계 정보를 활용
  - 효율적인 행렬 인수분해 방식
  - 의미 관계를 벡터 연산으로 표현 가능
- 응용 분야 : 단어 유사도 측정, 문서 분류, 기계 번역 등

2. Challenges

- 기존 방법(Word2Vec, LSA)의 한계 :
  - Word2Vec: local context만 사용
  - LSA : 계산 비용이 크고 희소 데이터 문제 존재
- 통계 정보를 효과적으로 활용하면서 계산 효율성 유지하는 것이 어려움

3. Method

- Co-occurrence Matrix : 전체 말뭉치에서 단어 i와 j의 공동 출현 횟수를 계산
- Cost Function : 공동 출현 확률 비율을 보존하도록 설계
- Weighted Least Squares 방식 적용
- 최종 벡터 : 의미 관계를 반영하는 임베딩 공간 형성
- 계산 효율성을 위해 전처리 단계에서 Co-occurrence Matrix를 미리 계산

4. Experiments

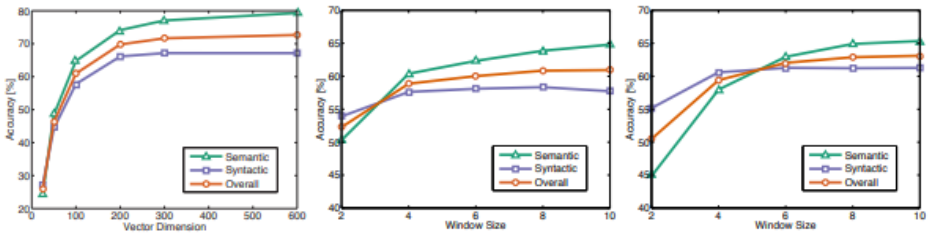
- 데이터셋 : Wikipedia 2014 + Gigaword 5 (6B tokens), Common Crawl (42B tokens)
- 평가 : 단어 유사도, 단어 관계 추론(Analogy task)
- 비교 baseline : Word2Vec, LSA
- 다양한 차원 수 및 window size 실험

5. Results

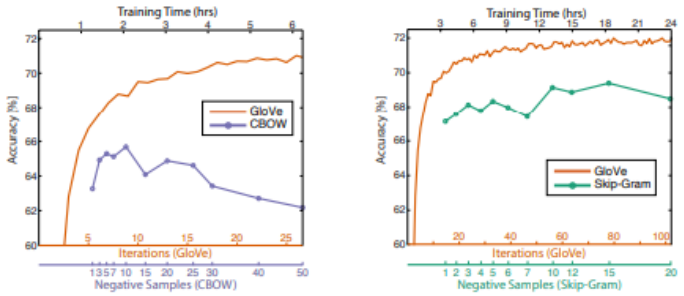
Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	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**

Model	Size	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	72.7	75.1	56.5	37.0
CBOW†	6B	57.2	65.6	68.2	57.0	32.5
SG†	6B	62.8	65.2	69.7	58.1	37.2
GloVe	6B	65.8	72.7	77.8	53.9	38.1
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	**75.9**	**83.6**	**82.9**	**59.6**	**47.8**
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

Model	Dev	Test	ACE	MUC7
Discrete	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
SVD-S	91.0	85.5	77.6	74.3
SVD-L	90.5	84.8	73.6	71.5
HPCA	92.6	**88.7**	81.7	80.7
HSMN	90.5	85.7	78.7	74.7
CW	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	**93.2**	88.3	**82.9**	**82.2**



(a) Symmetric context (b) Symmetric context (c) Asymmetric context



(a) GloVe vs CBOW (b) GloVe vs Skip-Gram

- GloVe는 단어 관계 추론(예: king − man + woman ≈ queen)에서 뛰어난 성능
- Word2Vec 대비 의미 관계 표현 능력이 향상됨
- 대규모 데이터셋에서 안정적인 벡터 학습 가능
- Co-occurrence 기반 접근이 local context 기반 방법보다 의미 보존에 강점 있음

6. Insight

- GloVe는 global co-occurrence 정보 활용이라는 점에서 Word2Vec와 차별화
- 단어 임베딩의 새로운 표준이 됨
- 후속 연구(FastText, contextual embeddings, Transformer 기반 임베딩)에 큰 영향