Sent2Vec 알아보기

오늘 공부할 논문

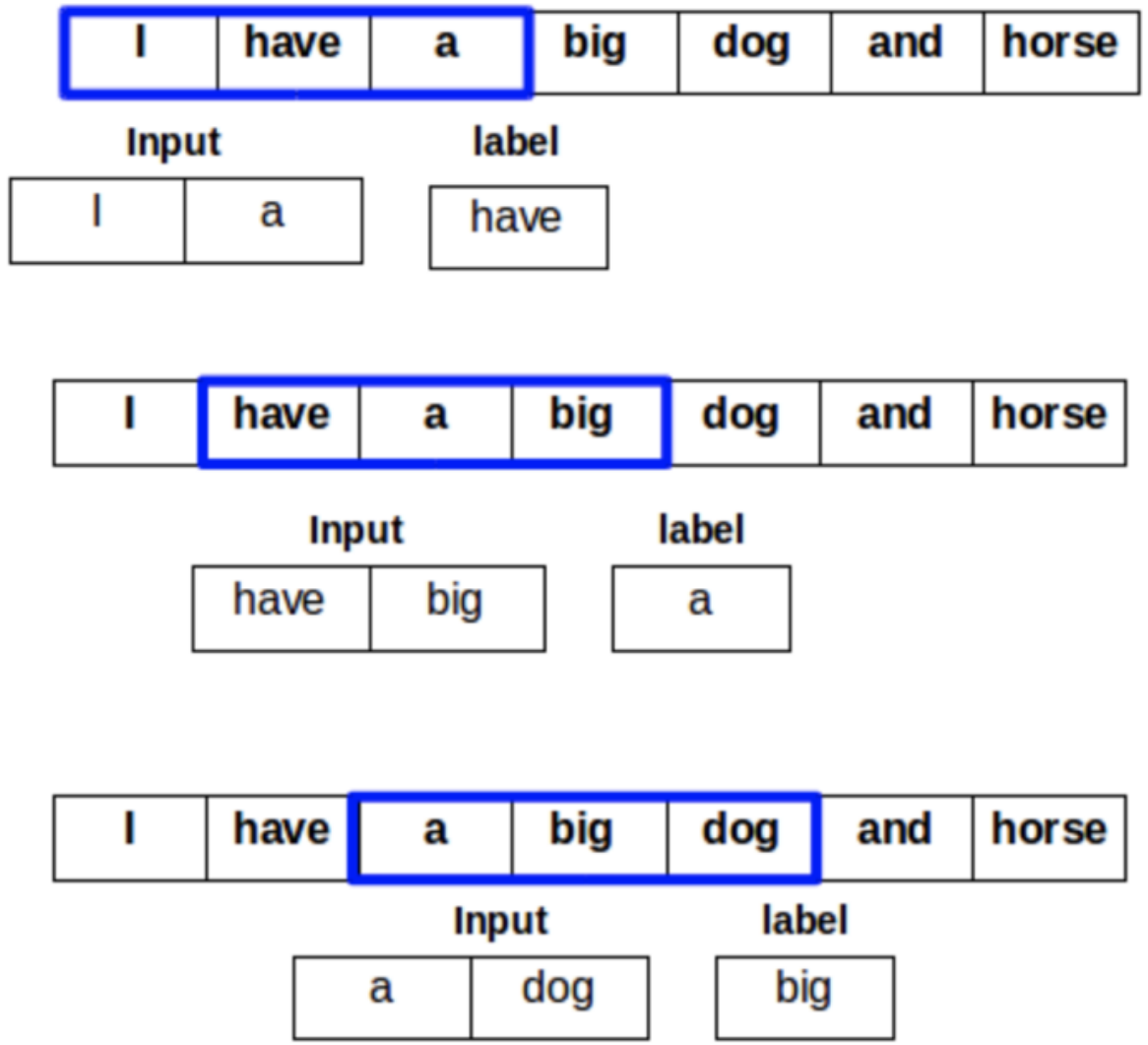
- Sent2Vec 문장 임베딩을 통한 한국어 유사 문장 판별 구현
  - 논문 링크: http://docs.likejazz.com/paper/kaonpark-sent2vec-2018.pdf
  - 요약: sent2vec을 이용한 문장 임베딩 기법으로 유사 문장 판별 시스템 구현. 한국어 특성에 맞게 모델을 변형하여 성능 향상

Sent2vec의 특징

- 비지도 학습이다.
- 벡터 하나가 문장 하나를 나타낸다.
- 코사인 유사도를 활용하여 문장의 유사도를 계산할 수 있다.
- 학습시 CBOW를 변형한 기법을 사용한다.

CBOW

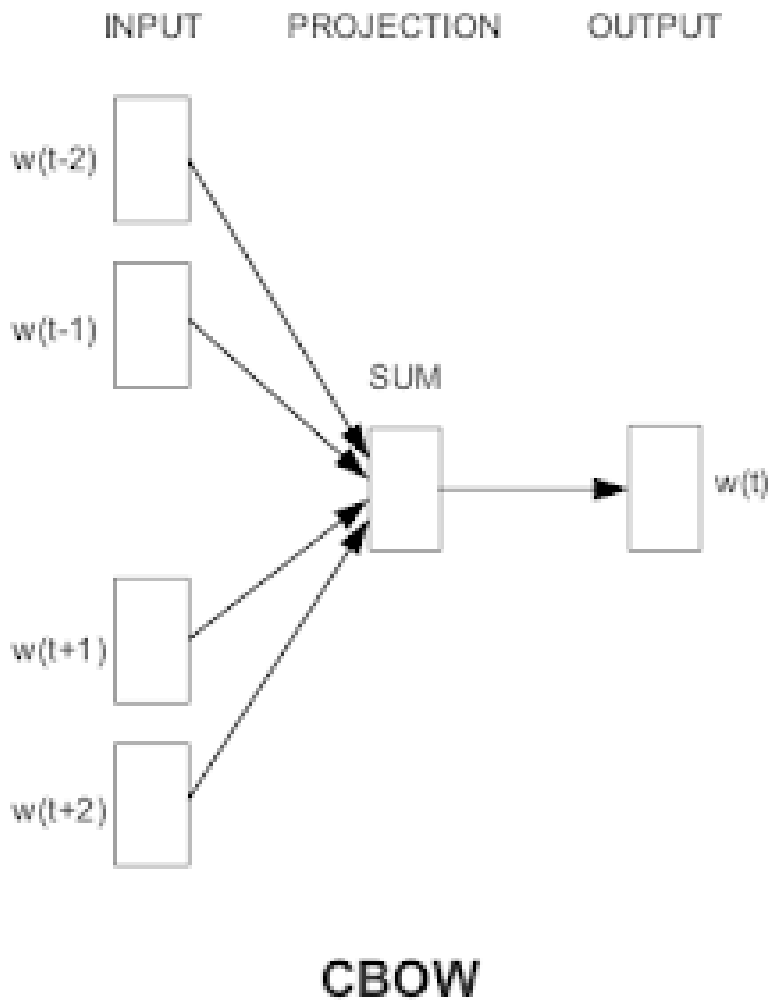
- CBOW는 학습할 때 window 내 중심 단어를 하나 둔다.
- 나머지 단어는 위치와 상관 없이 전부 중심 단어 학습에 사용된다.
- CBOW는 문맥, 즉 중심 단어 주변에 있는 k개의 단어를 보면 중심 단어를 유추할 수 있도록 학습한다.



http://zongsoftwarenote.blogspot.com/2017/04/word2vec-model-introduction-skip-gram.html

CBOW

- 우측 그림의 화살표는 행렬을 나타낸다.
- Input-Projection 간의 행렬은 모든 단어가 공통이다.
- 행렬 값을 구할 때 신경망 기법 사용.
- 모델의 Input/Output은 단어를 one-hot encoding으로 하여 학습 수행



CBOW

http://zongsoftwarenote.blogspot.com/2017/04/word2vec-model-introduction-skip-gram.html

Sent2Vec의 CBOW 활용

일반적인 CBOW

- Subsampling 사용
  - 학습 속도를 올리기 위해
- Window Size가 고정
  - 단어 임베딩이 목적이라서, 근처 단어 간의 상관관계를 보기 위해

Sent2Vec에서의 CBOW

- Window Size는 문장 전체 길이 (Dynamic Context Window)
  - 문장 전체의 n-gram 쌍을 보려고
- Subsampling 사용 X
  - 문맥 파악에 중요한 n-gram 쌍을 subsampling으로 날릴 수 있으므로

Sent2Vec의 학습 방법

- 변형된 CBOW를 이용하여, n-gram 쌍 (unigram 포함)에 대하여 학습한다.
  - n-gram 쌍은 일반적인 의미와 달리, 2-gram 쌍이되 뒤쪽으로 확장된 의미이다.
- 예시 문장: 한국어 자연어 처리는 힘들다
  - 현재 단어: 한국어
  - 단어 n-gram=2: (한국어), (한국어 자연어)
  - 단어 n-gram=3: (한국어), (한국어 자연어), (한국어 처리는)
  - …

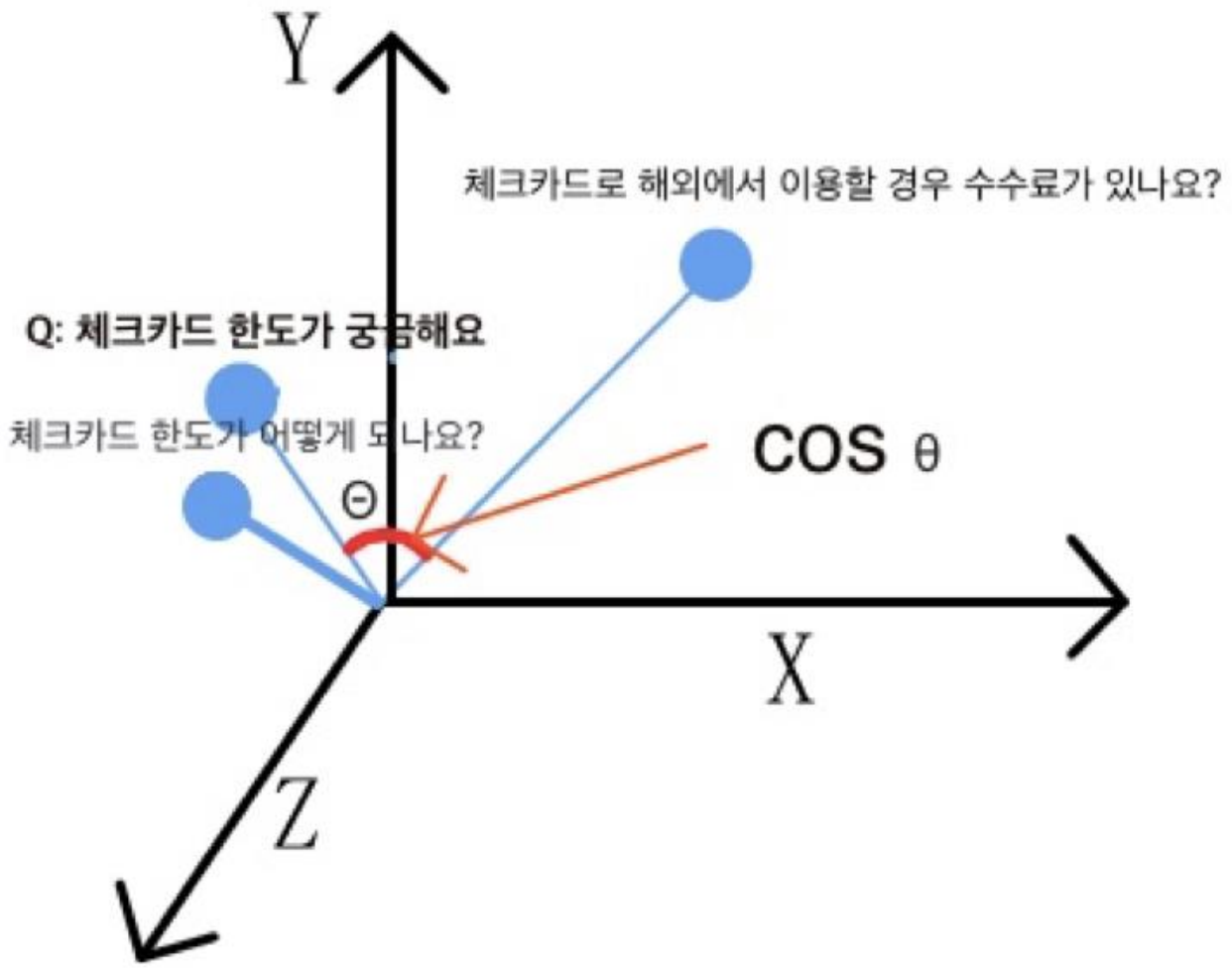
Sent2Vec 값 구하기

- 앞서 학습 결과, 모든 단어 n-gram 쌍의 벡터값을 확인할 수 있다.
- 단순히 문장에 있는 모든 단어 n-gram 쌍(unigram 포함)의 벡터값의 평균을 가지는 값을 문장 임베딩으로 갖는다.
- 문서 내 모든 단어의 Word2Vec 값을 평균낸 것으로 봐도 무방.

$$\boldsymbol{v}_S := \frac{1}{|R(S)|}\boldsymbol{V}\iota_{R(S)} = \frac{1}{|R(S)|}\sum_{w\in R(S)}\boldsymbol{v}_w$$

문장 유사도 구하기

각 문장의 sent2vec의 값을 코사인 유사도를 구하면 두 문장의 유사도를 구할 수 있다.



한국어 특성에 맞게 변형

- 주어부 가중치
  - 우리 말은 주어부에 중요한 단어가 등장한다는 가정 하에, 문장의 절반 지점 앞을 주어부로 보고 가중치 적용. 가중치는 1.2~1.5 사이.
- 코퍼스 출현 빈도에 따른 가중치 감소
  - 이 논문에선 t = 1000

$$w = \frac{(t+1) \cdot k}{t+k}$$

실험

구현

- C++로 구현된 fasttext 라이브러리에 sent2vec 알고리즘 추가 및 개선

학습 데이터

- 2013~2015년의 모든 언론사 뉴스를 형태소 분석한 결과
- 4.5억개 문장, 100억개 단어, 용량은 85GB

테스트 데이터

- 실험자가 직접 5000개 문장으로 구성된 테스트셋 생성

실험 결과

Table 2: 각 모델에 따른 성능 평가 결과

모델	P@1	P@3	P@5
Word2Vec CBOW	0.7574	0.8618	0.8927
Sent2Vec uni. (baseline)	0.7854	0.8997	0.9271
Sent2Vec bi.	0.8306	0.9296	0.9539
Sent2Vec bi. + 주어부 가중치 1.2	0.8369	0.9312	0.9544
Sent2Vec bi. + 주어부 가중치 1.2 + 가중치 감소*	**0.8499**	**0.9374**	**0.9597**

요약

- Sent2Vec은 문장 하나를 하나의 벡터로 임베딩하는 것이다
- window size를 문장 전체로 늘리는 등, CBOW를 변형한 방식으로 학습한다.
- 문장 내 모든 n-gram 쌍의 벡터값을 평균하면 문장 임베딩 값을 구할 수 있다.
- Sent2vec 값을 코사인 유사도 취하면 문장의 유사도를 구할 수 있다.

Reference

- http://docs.likejazz.com/paper/kaonpark-sent2vec-2018.pdf
- http://docs.likejazz.com/sent2vec/#sent2vec
- https://arxiv.org/abs/1703.02507
- http://zongsoftwarenote.blogspot.com/2017/04/word2vec-model-introduction-skip-gram.html
- https://datascienceschool.net/view-notebook/6927b0906f884a67b0da9310d3a581ee/