Brain neuro
Language
Processing

Bilingual Word Embedding with parallel corpus

고려대학교 컴퓨터학과
Brain-neuro Language Processing Lab
이설화
E-mail : whiteldark@korea.ac.kr

Index

Introduction to Word Representation

Trend of word representation

- 단어를 dense 실수 벡터인 vector representation으로 표현하는 방법(word embedding)은 최근 NLP분야에서 인기를 얻고 있음

dog

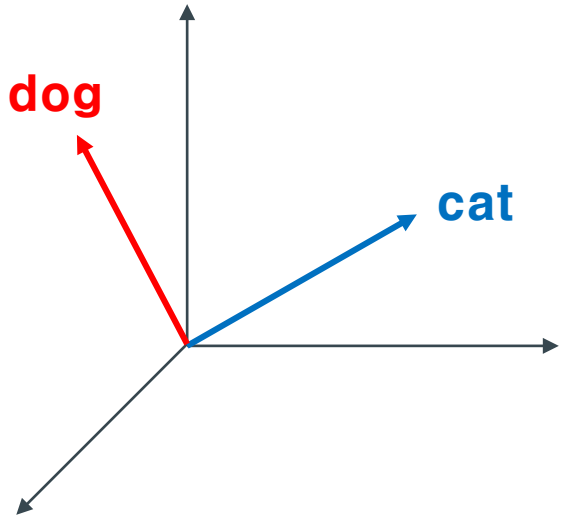
cat

Dense vector

0.3	0.8	0.9	0.1	0.3

0.5	0.1	0.1	0.2	0.6

Vector representation



Traditional word representation

- 기존 word representation은 one-hot (or one-of-N) encoding방식을 사용해왔음

Vocabulary (size=10)

word	id
“the”	1
“dog”	2
“cat”	3
“and”	4
…	…

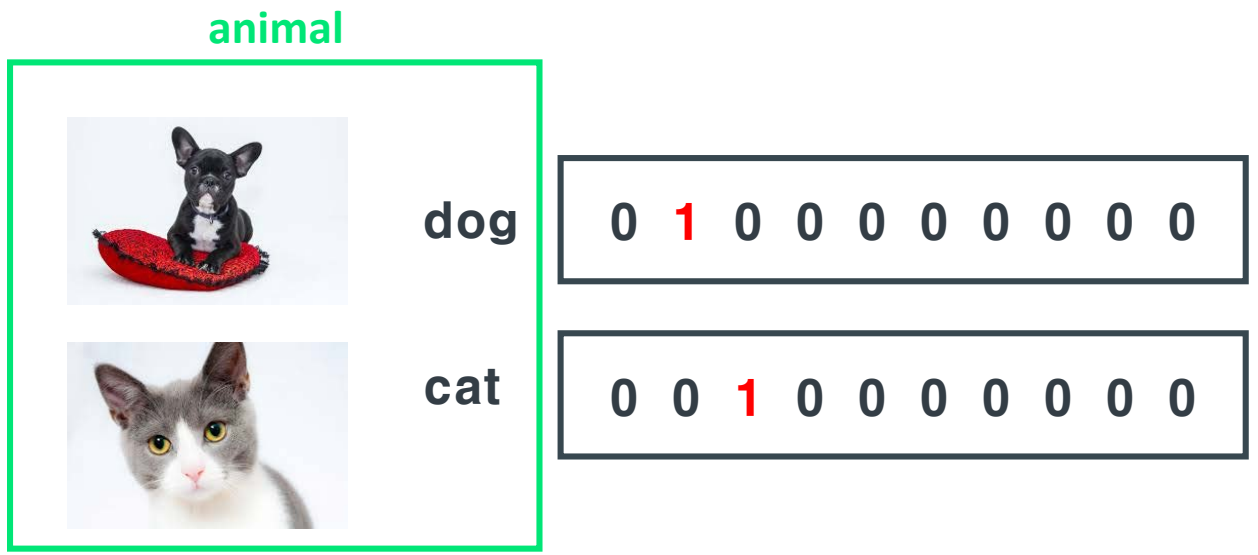
One-hot vector representation

dog 0 1 0 0 0 0 0 0 0 0

cat 0 0 1 0 0 0 0 0 0 0

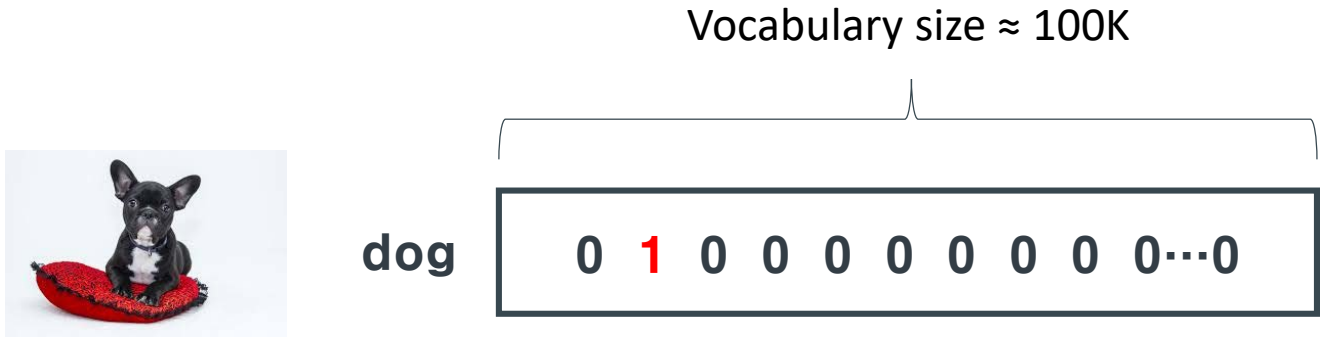
Limitation of traditional word representation method (1/2)

- One-hot encoding은 단어들 간의 관계성을 고려하여 표현하지 않음
  - Ex. Cat : Id3, dog : Id2 => animal



Limitation of traditional word representation method (2/2)

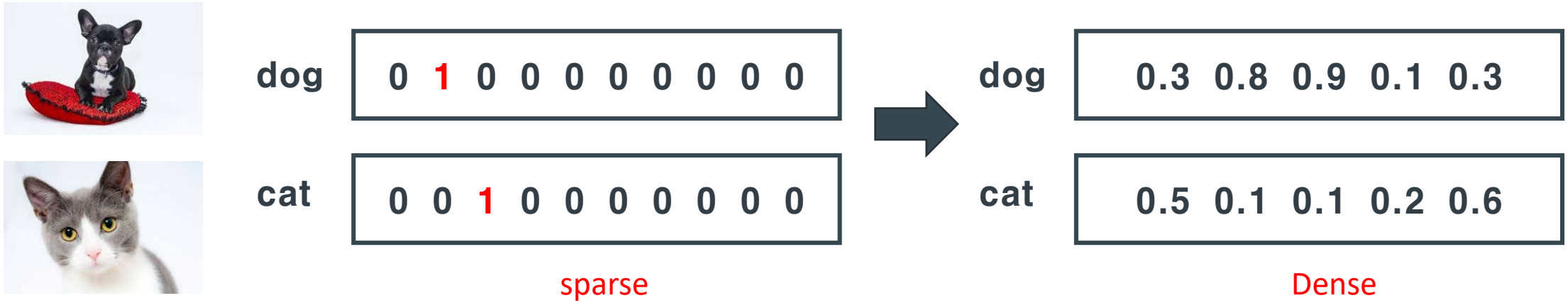
- One-hot representation은 매우 높은 dimension을 가지는 문제점이 있음
  - Memory expensive



Word Embedding

Word embedding이란?

- 일반적으로 one-hot encoding과 같은 sparse한 방식으로 표현된 단어를 더 낮은 차원의 dense한 실수 벡터 공간에 매핑하는 것
  - 한 단어당 일차원 dimension의 벡터 공간으로 매핑함으로써 lower dimension을 구성함

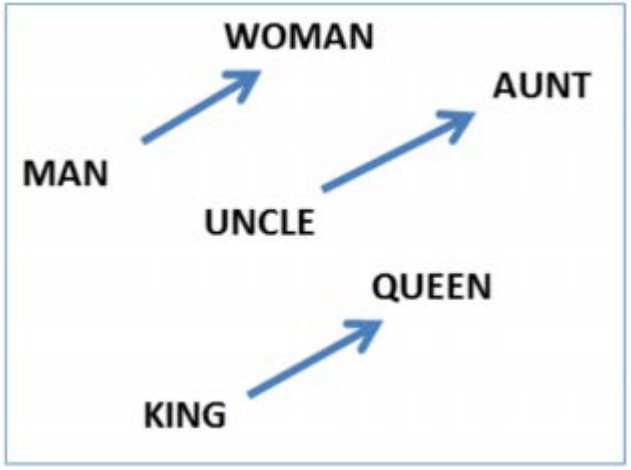


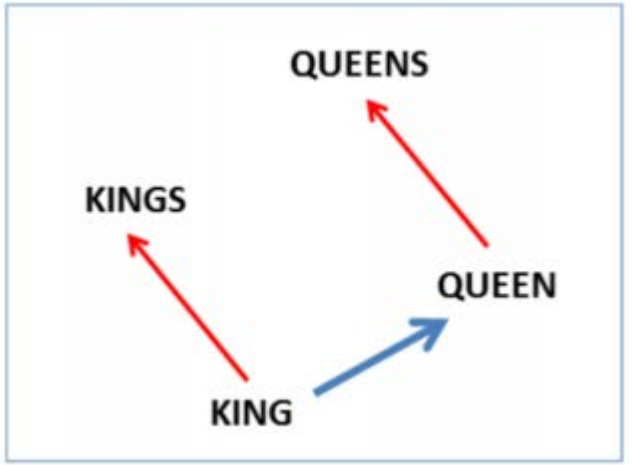
Word embedding의 목표

- 단어를 dense한 실수 벡터 공간에 매핑하되,
  단어의 의미가 반영되도록 함
  - 유사한 의미의 단어는 벡터 공간 상의 가까운 거리 내에 분포하도록 함

Word embedding의 특성

- 단어 의미 간의 상관관계도 반영됨
- $king - man + woman \approx ???$
- $king - man + woman \approx queen$
- Example 사이트 : http://w.elnn.kr/search/





Bilingual Word Embedding

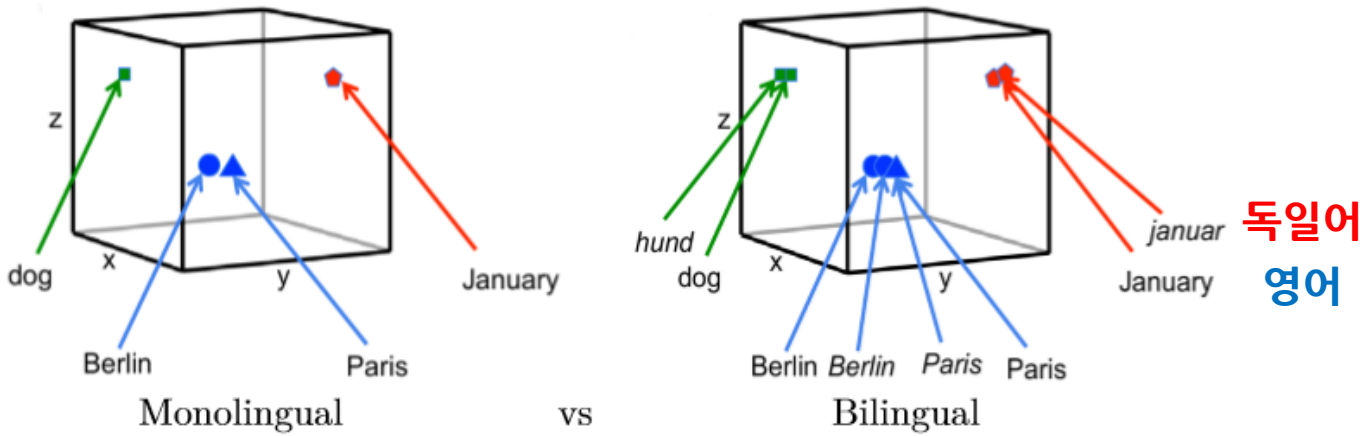
Bilingual Word Embedding 정의

Bilingual Word Embedding (BWE) 정의

- 두 개의 다른 언어로부터 하나의 shared space에 단어를 embedding하는 것

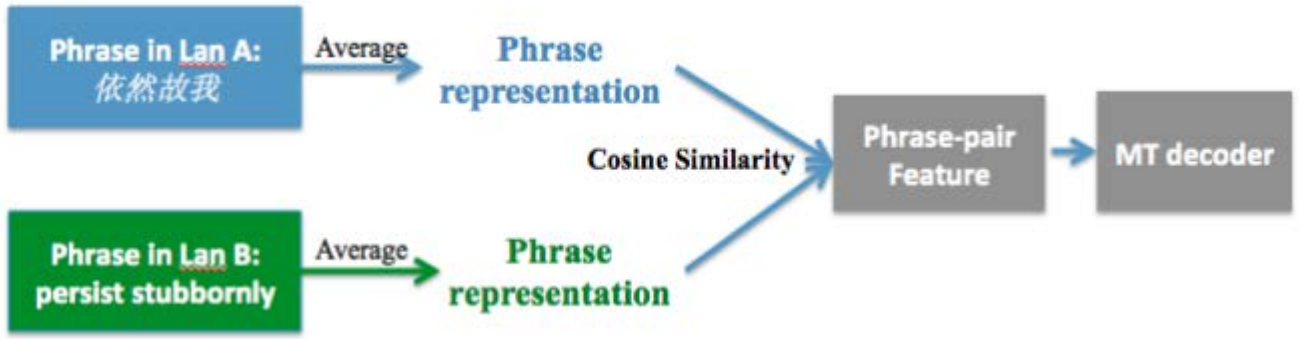
BWE goal

- BWE learning model은 서로 다른 두 언어에서 유사한 의미를 가지는 단어가 유사한 공간에 mapping되도록 하는 것



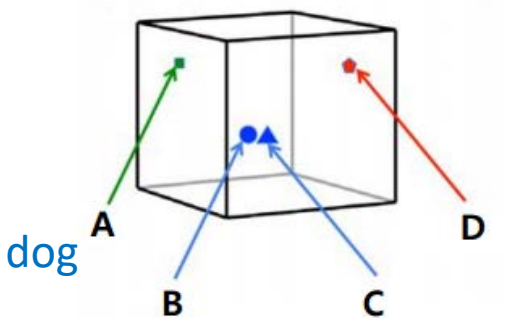
Uses of BWE

- Machine translation (Zou et al., 2013)



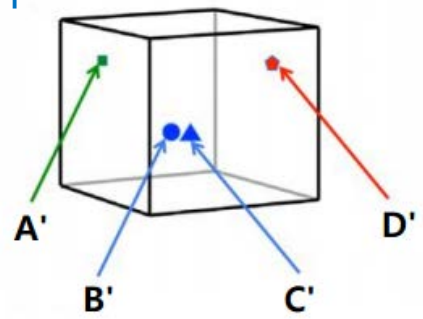
Overview of building a BWE space

영어



Source WES

한국어



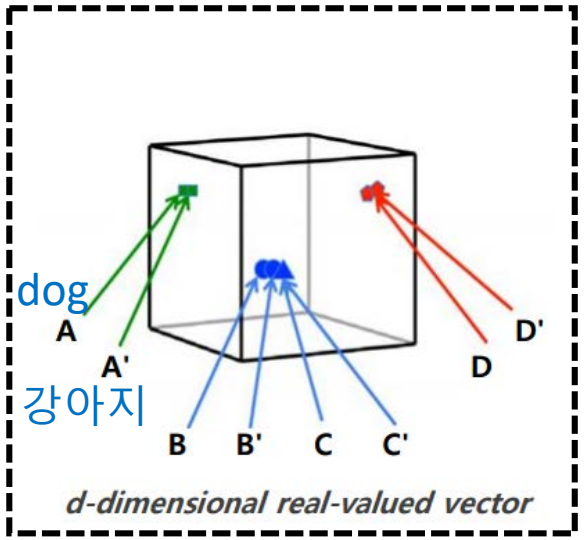
Target WES

$V^S = A, B, C, D$

$V^T = A', B', C', D'$

Using seed lexicon

Shared BWE space



Mapping

유사한 의미의 단어를 유사한 공간에 mapping

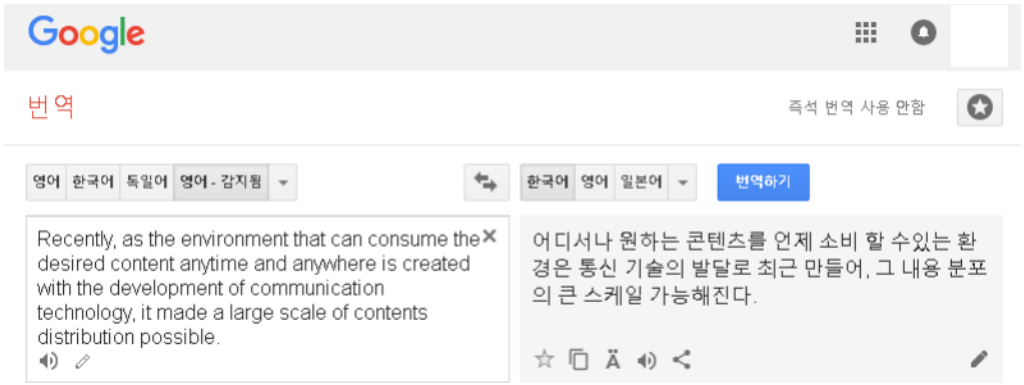
Building a Bilingual word embedding space with parallel corpus

Research Motivation (1/2)

- 기존의 BWE 연구에서는 **“High-quality training seed lexicons”**가 존재한다고 가정
  - Vulic, Ivan, and Anna Korhonen. "On the role of seed lexicons in learning bilingual word embeddings." ACL, 2016.
  - Shared Bilingual Word Embedding Space(SBWES)를 유도하는 과정에서 seed lexicons 의 역할과 중요성에 대해 분석하였음

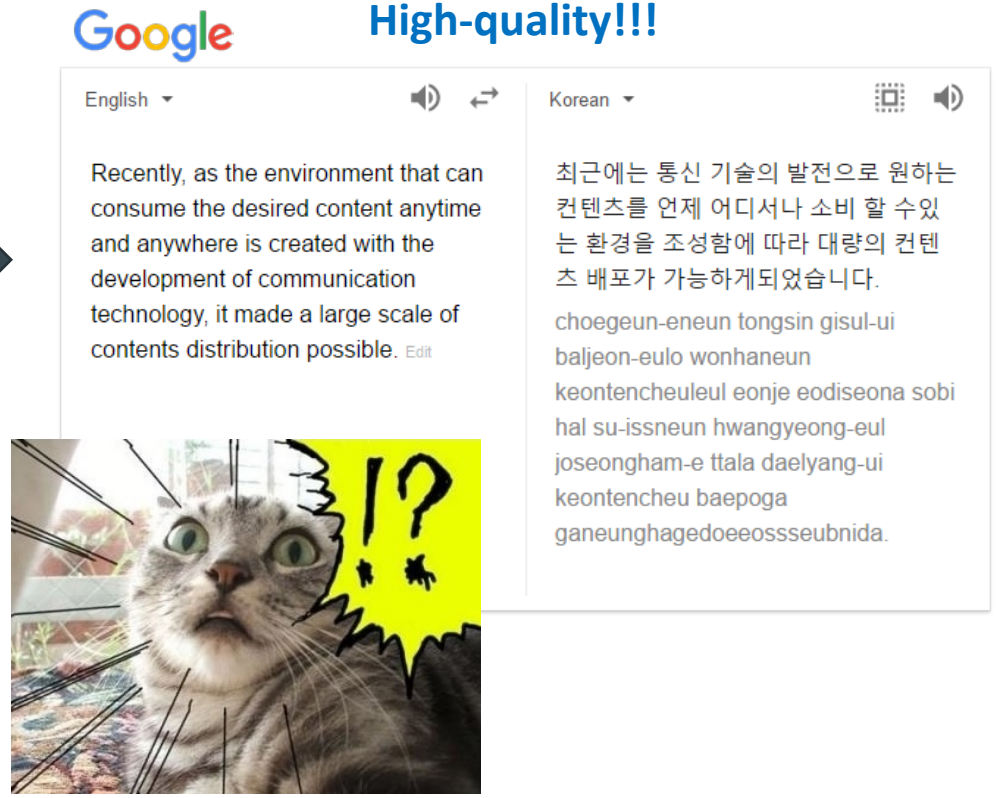
Research Motivation (1/2)

High-quality???



upgrade

High-quality!!!



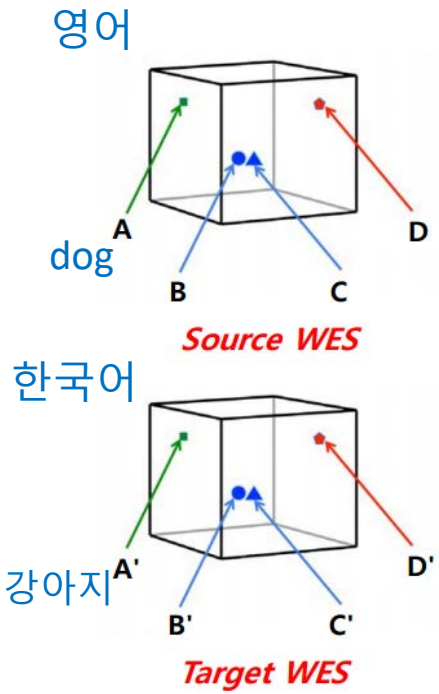
Research Motivation (2/2)

SBWES를 생성하는 데는 다음의 두 가지 과정을 거침 (Vulic et al.,2016,ACL)

1

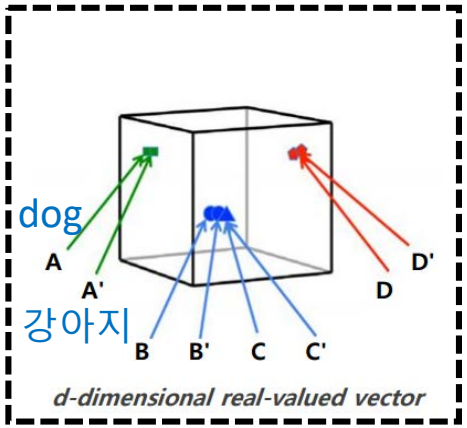
두 개의 별도로 준비 된 non-aligned monolingual embedding spaces는 monolingual WE learning model을 사용하여 유도하게 됨

- 비교적 얻기 쉬운 문서 번역쌍 (Wikipedia)을 word translation pairs로 사용하였음
- Our research idea => Parallel corpus



Using seed lexicon

1

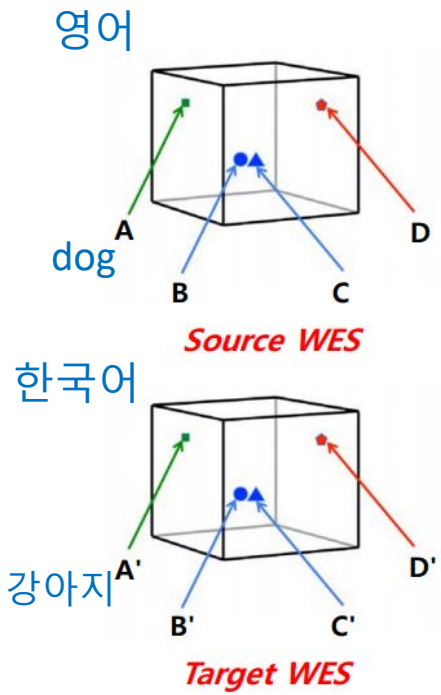


Mapping

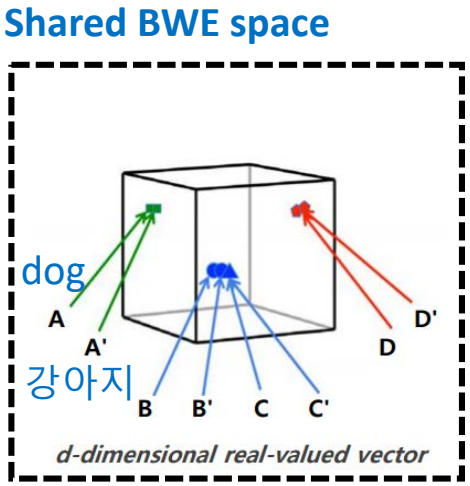
Research Motivation (2/2)

2

단어 번역 쌍을 seed lexicon으로 사용하고, 이를 통해 두 개의 monolingual spaces를 하나로 묶는 mapping function을 학습함으로써 shared BWE space를 구성함



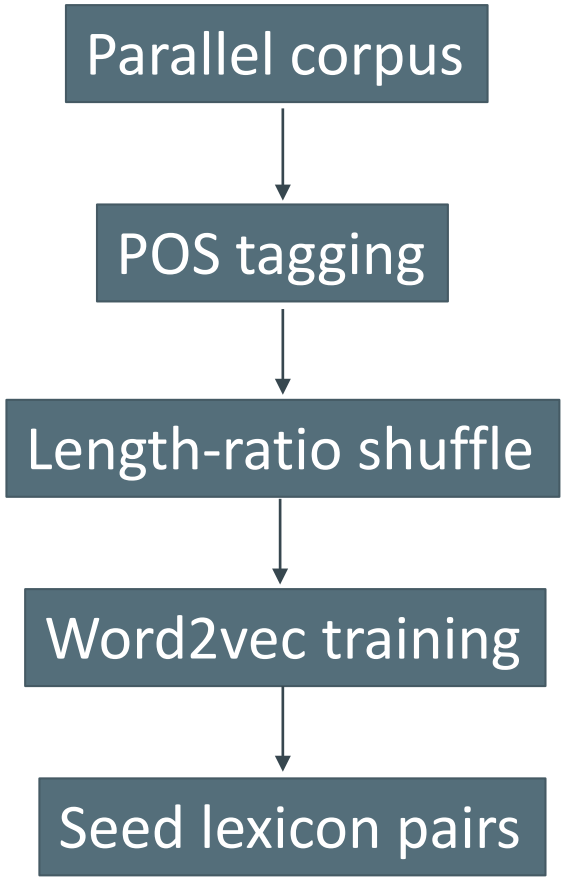
Using seed lexicon



Mapping

2

making the seed lexicons



Making the seed lexicons

- Parallel corpus
  - Sentence-aligned corpus
  - Bible data (총 66권, 1,189장, 31,102 절로 구성)

Making the seed lexicons

권, 장, 절

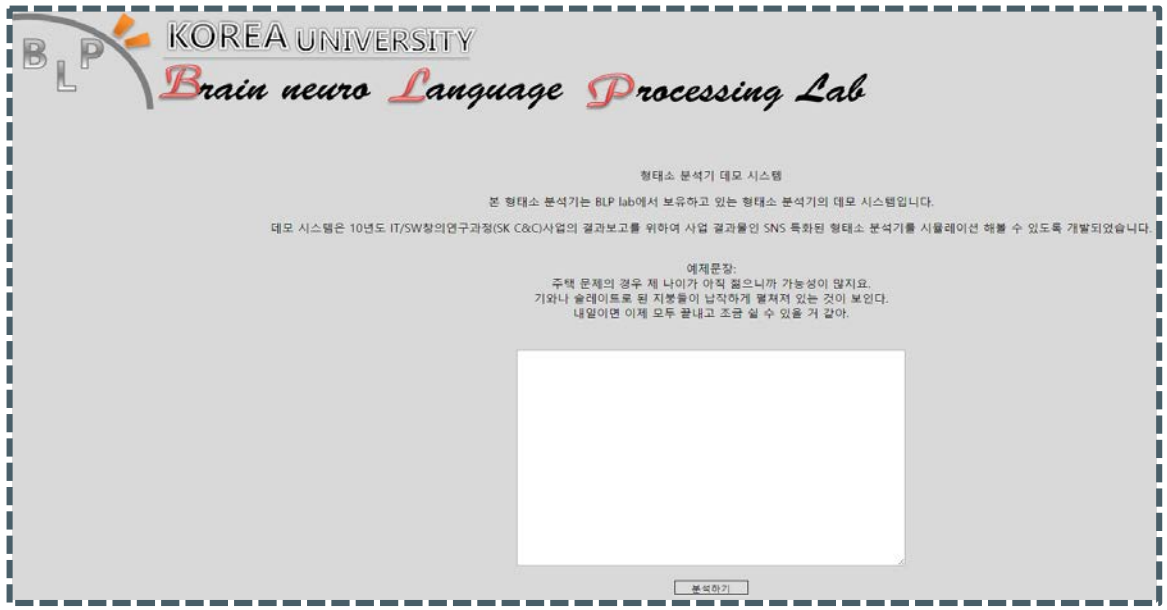
```
01창 1:1 태초에 하나님이 천지를 창조하시니라
01창 1:2 땅이 혼돈하고 공허하며 흑암이 깊음 위에 있고 하나님의 영은 수면 위에 운행하시니라
01창 1:3 하나님이 이르시되 빛이 있으라 하시니 빛이 있었고
01창 1:4 빛이 하나님이 보시기에 좋았더라 하나님이 빛과 어둠을 나누사
01창 1:5 하나님이 빛을 낮이라 부르시고 어둠을 밤이라 부르시니라 저녁이 되고 아침이 되니 이는 첫째 날이니라
01창 1:6 하나님이 이르시되 물 가운데에 궁창이 있어 물과 물로 나뉘라 하시고
01창 1:7 하나님이 궁창을 만드사 궁창 아래의 물과 궁창 위의 물로 나뉘게 하시니 그대로 되니라
01창 1:8 하나님이 궁창을 하늘이라 부르시니라 저녁이 되고 아침이 되니 이는 둘째 날이니라
01창 1:9 하나님이 이르시되 천하의 물이 한 곳으로 모이고 뭍이 드러나라 하시니 그대로 되니라
```

한국어

```
01Gn 1:1 In the beginning God created the heavens and the earth.
01Gn 1:2 Now the earth was formless and empty, darkness was over the surface of the deep, and the Spirit of God was hovering over the waters.
01Gn 1:3 And God said, "Let there be light," and there was light.
01Gn 1:4 God saw that the light was good, and he separated the light from the darkness.
01Gn 1:5 God called the light "day," and the darkness he called "night." And there was evening, and there was morning--the first day.
01Gn 1:6 And God said, "Let there be an expanse between the waters to separate water from water."
01Gn 1:7 So God made the expanse and separated the water under the expanse from the water above it. And it was so.
01Gn 1:8 God called the expanse "sky." And there was evening, and there was morning--the second day.
01Gn 1:9 And God said, "Let the water under the sky be gathered to one place, and let dry ground appear." And it was so.
```

영어

Making the seed lexicons
Preprocessing

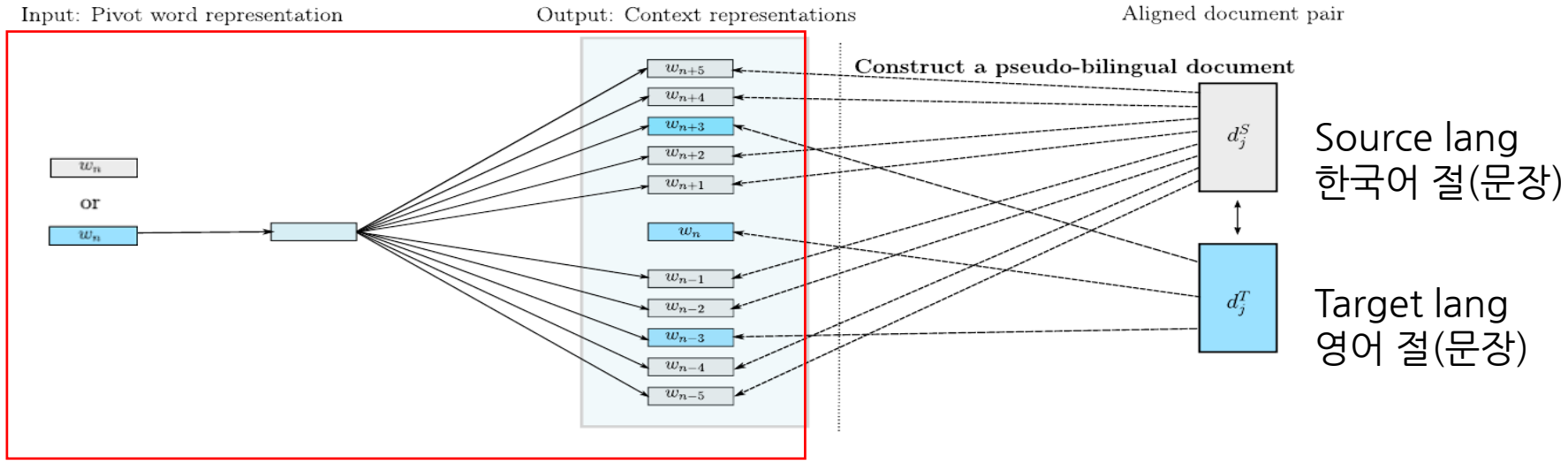


한국어 형태소 분석기

- 기호 제거
- 한국어
  - POS tagger (BLP lab asset)
  - url : http://blpdemo.korea.ac.kr/MA/

Making the seed lexicons
Preprocessing

- 각 문장별 token수의 비율로 shuffle
  - 예) Ko = {포도, 복숭아}
    - Token number = 2
  - 예) Eng = {carrot, apple, pineapple, egg}
    - Token number = 4
    - Length-ratio = 2 : 1



SGNS

Length-ratio shuffle

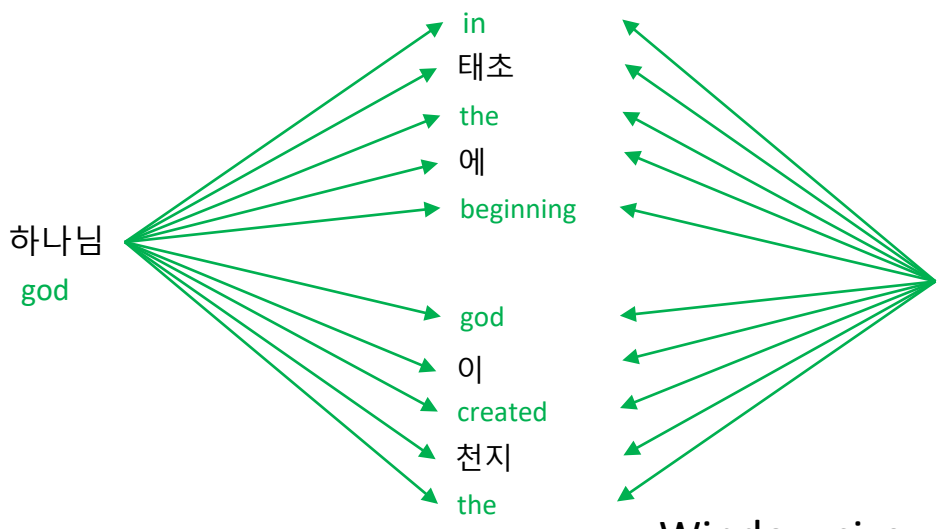
Making the seed lexicons
Bilingual Word Embedding Skip-Gram (1/2)


```
01창 1:1 태초에 하나님이 천지를 창조하시니라
01창 1:2 땅이 혼돈하고 공허하며 흑암이 깊음 위에 있고 하나님의
01창 1:3 하나님이 이르시되 빛이 있으라 하시니 빛이 있었고
01창 1:4 빛이 하나님이 보시기에 좋았더라 하나님이 빛과 어둠을
01창 1:5 하나님이 빛을 낮이라 부르시고 어둠을 밤이라 부르시니라
01창 1:6 하나님이 이르시되 물 가운데에 궁창이 있어 물과 물로
01창 1:7 하나님이 궁창을 만드사 궁창 아래의 물과 궁창 위의 물
01창 1:8 하나님이 궁창을 하늘이라 부르시니라 저녁이 되고 아침이
```

Korean

```
01Gn 1:1 In the beginning God created the heaven
01Gn 1:2 Now the earth was formless and empty,
01Gn 1:3 And God said, "Let there be light," and
01Gn 1:4 God saw that the light was good, and he
01Gn 1:5 God called the light "day," and the dar
01Gn 1:6 And God said, "Let there be an expanse
01Gn 1:7 So God made the expanse and separated
01Gn 1:8 God called the expanse "sky." And there
01Gn 1:9 And God said, "Let the water under the
```

English



Window size = 5

```
in 태초 the 에 beginning 하나님 god 이 created 천지 the 를 heavens 창조 and 하 the earth 시
땅 now 이 the 혼돈 earth 하 was 고 formless 공허하 and 며 empty 흑암 darkness 이 was 깊 over
하나님 and 이 god 이르 시 said 되 let 빛 이 there 있 be 으라 하 light 시 and 니 빛 there 이 was
빛 god 이 saw 하나님 that 이 the 보 light 시 기 was 에 good 좋 and 았 he 더라 separated 하나님
하나님 god 이 called 빛 을 the 낮 light 이 라 day 부르 and 시 the 고 어둠 darkness 을 he 밤 이
하나님 and 이 god 이르 said 시 되 let 물 there 가운데 be 에 an 궁창 이 expanse 있 between 어 th
하나님 so 이 god 궁창 made 을 the 만들 expanse 사 and 궁창 separated 아래 the 의 water 물 unde
하나님 god 이 called 궁창 을 the 하늘 expanse 이 라 sky 부르 and 시 니라 there 저녁 was 이 even
하나님 and 이 god 이르 said 시 let 되 the 천하 water 의 under 물 the 이 sky 한 be 곳 gathered
```

Length-ratio shuffle

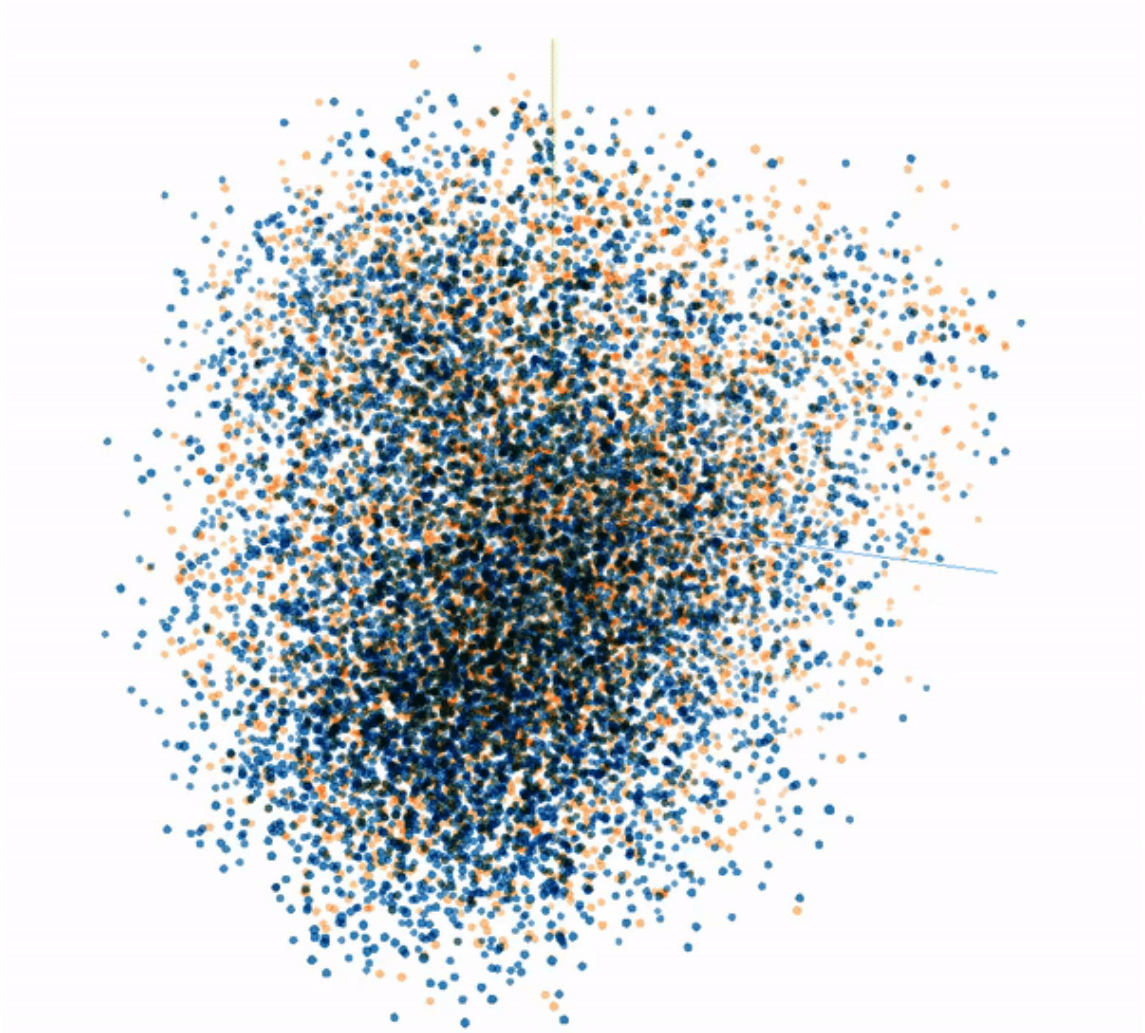
Making the seed lexicons
Bilingual Word Embedding Skip-Gram (2/2)

- Word2vec package
  - SGNS (Skip-gram Negative Sampling)

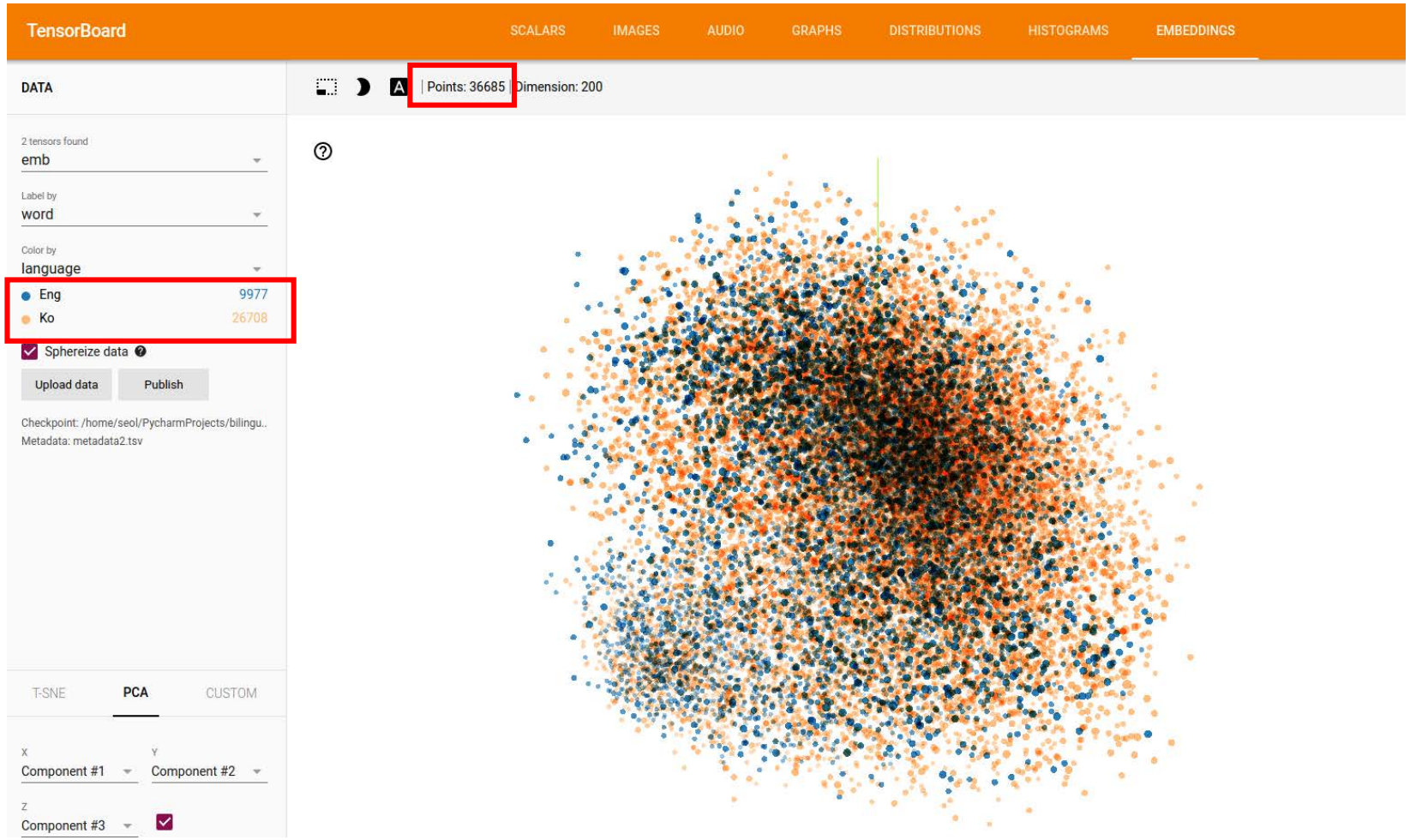
Visualization of seed lexicons

Google

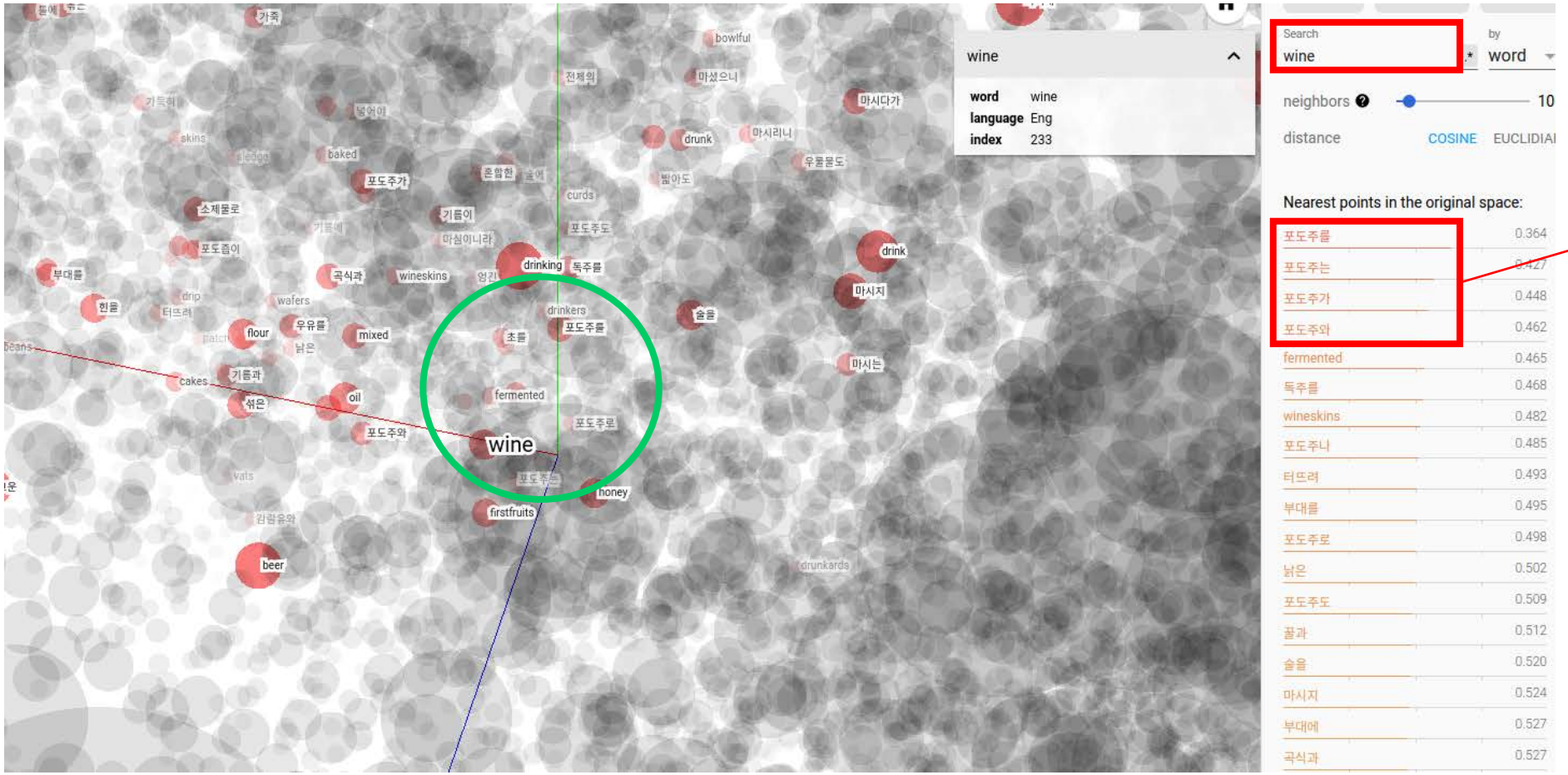
TensorBoard



Visualization of seed lexicons_어절단위

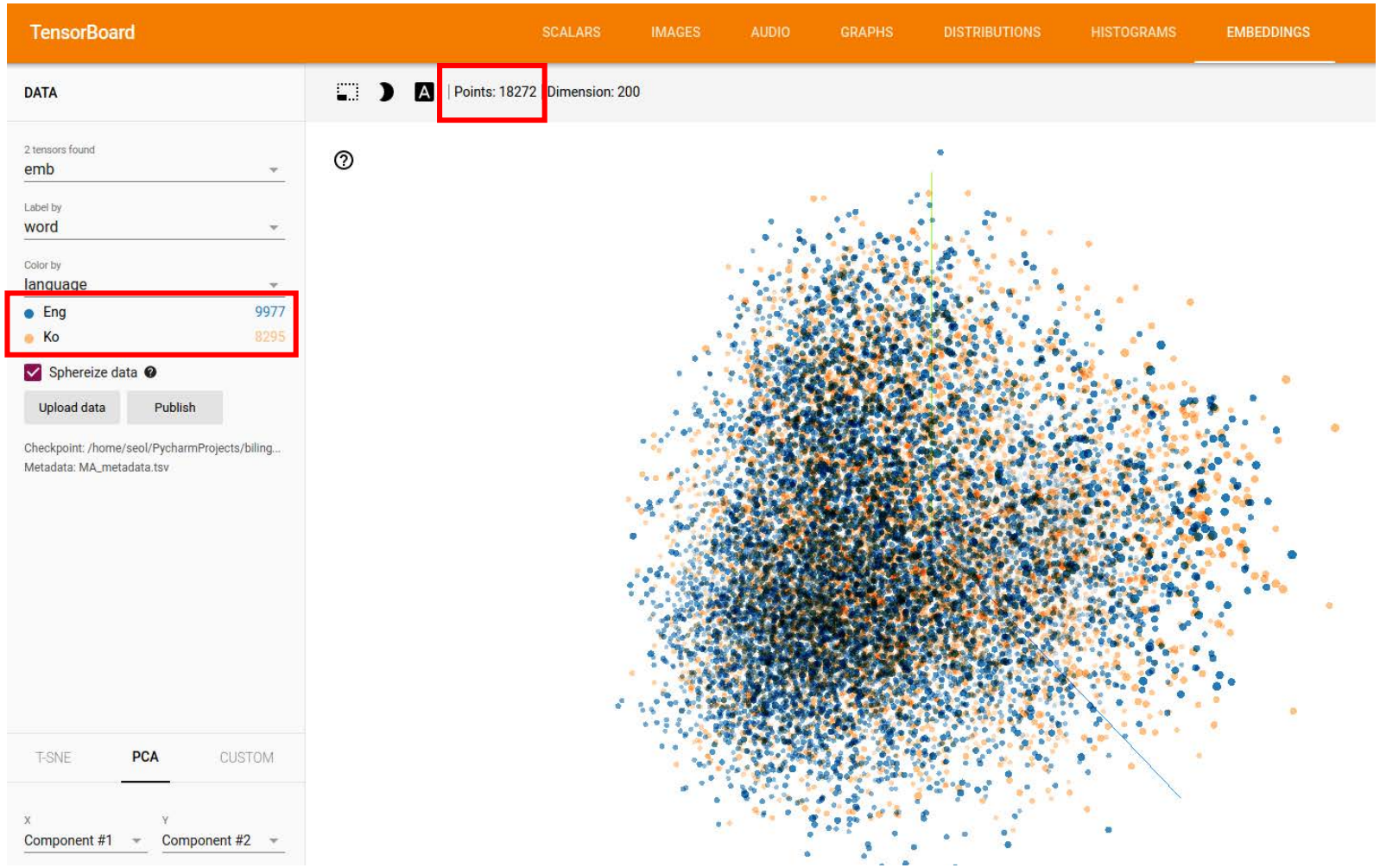


Visualization of seed lexicons_어절단위



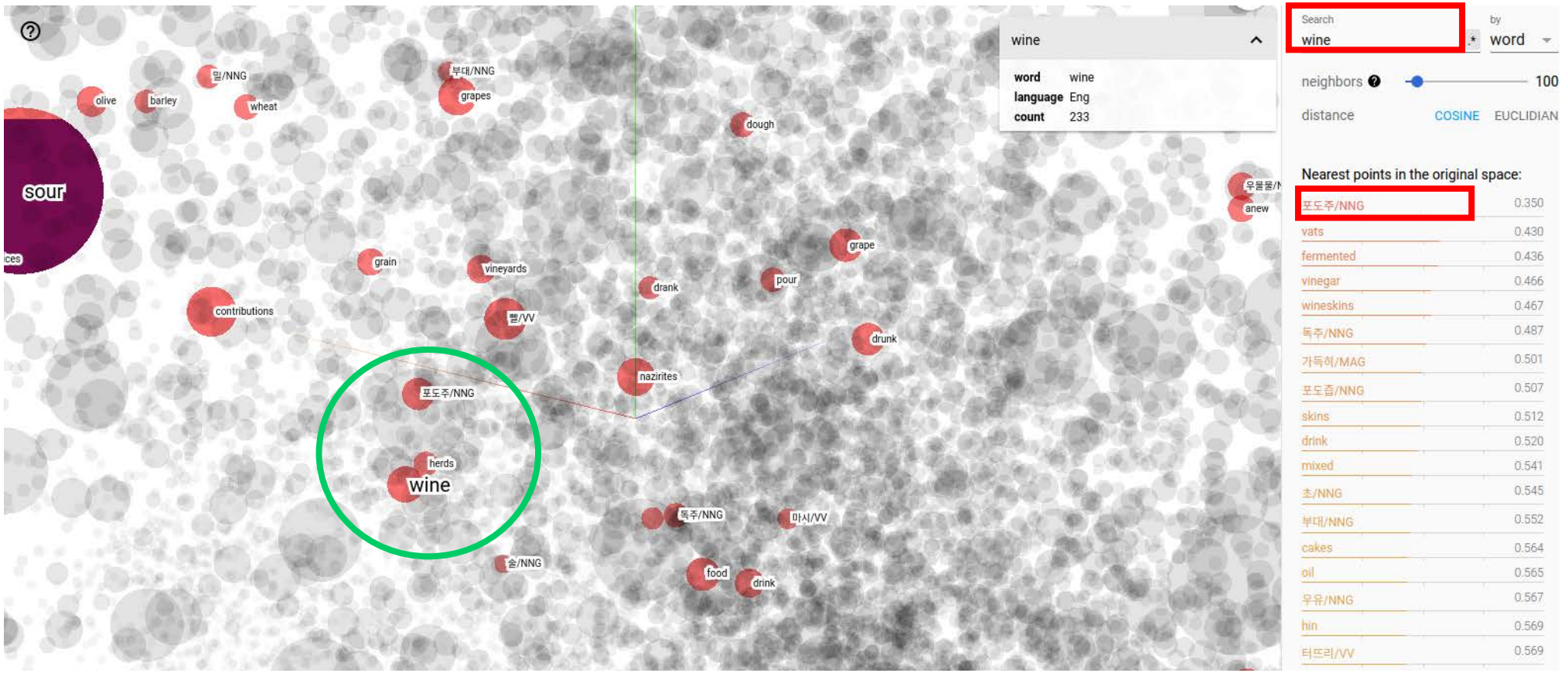
어절 단위의 결과

Visualization of seed lexicons_형태소 단위

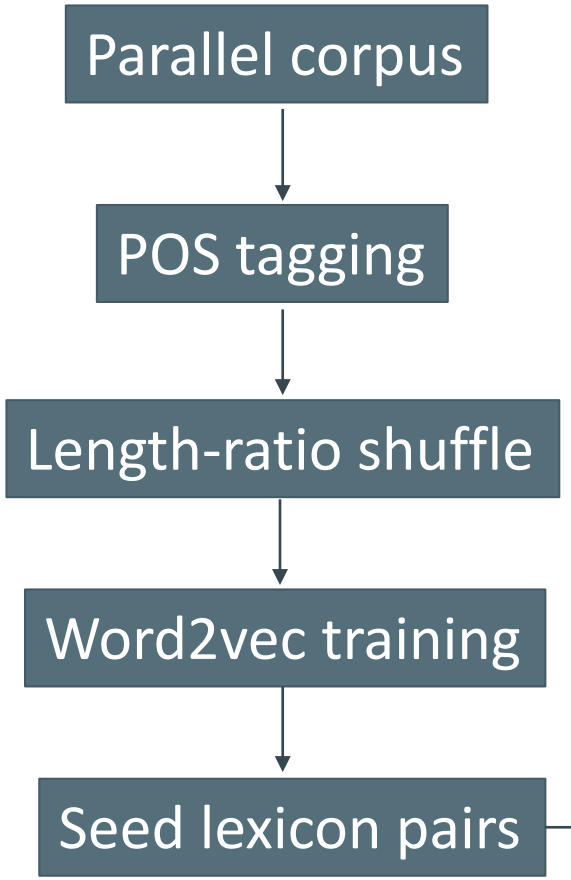


Visualization of seed lexicons _형태소 단위

"Wine" 검색 => "포도주"가 가장 유사한 결과로 나옴



Mapping function



L2-regularized least-squares error objective

$$\min_{\mathbf{W}\in\mathbb{R}^{d_S\times d_T}} \|\mathbf{XW}-\mathbf{Y}\|_F^2+\lambda\|\mathbf{W}\|_F^2$$

W *: 2차원 행렬*

Mapping matrix

X *: Source 언어의 word 벡터*

Y *: Target 언어의 word 벡터*

X, Y *: Seed lexicon*

Mapping function

Seed lexicon demo URL

http://blplab.iptime.org:6006

References

Vulic, Ivan, and Anna Korhonen. "On the role of seed lexicons in learning bilingual word embeddings." ACL, 2016.

Vulić, Ivan, and Marie-Francine Moens. "Bilingual distributed word representations from document-aligned comparable data." Journal of Artificial Intelligence Research 55 (2016): 953-994.

Zou, Will Y., et al. "Bilingual Word Embeddings for Phrase-Based Machine Translation." EMNLP. 2013.

감사합니다.

이설화
E-mail : whiteldark@korea.ac.kr