숫자로 표상된 의미:
기계학습 도구 Word2vec 사용기

최재웅(고려대)
언어정보학회
2018년 4월 14일

목 차

발표 순서

1. 도입: 의미 표상
2. 워드벡터/워드임베딩
   – 원리, 특징, 장점
   – Q: 어떻게 구하는가?
   – A: 빅데이터와 기계학습 기법(예: Word2vec) 활용
3. 결과자료
4. 언어학에서의 활용방안?

"Semantic Web"

- "I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers

The term was coined by <u>Tim Berners-Lee</u> for a web of data (or **data web**) that can be processed by machines—that is, one in which much of the meaning is machine-readable. (From Wikipedia)

도 입

What deep learning has achieved so far

François Chollet 2018

- Superhuman Go playing
- Near-human-level image classification
- Near-human-level speech recognition
- Near-human-level handwriting transcription
- Near-human-level autonomous driving
- Digital assistants such as Google Now and Amazon Alexa
- Ability to answer natural-language questions
- Improved machine translation
- Improved text-to-speech conversion
- Improved search results on the web
- Improved ad targeting, as used by Google, Baidu, and Bing

의미표상

- Dictionary definition
  - Natural languages: 엄마 ➔ ‘female parent’
- Componential analysis
  - Meaning components: 엄마 ➔ [+female, ↑ parent]
- Formal semantics
  - Model theoretic mapping: 엄마 ➔ $[\![\text{mother}]\!]^{\mathcal{M},g}$
- Distributional hypothesis, ‘quantitative turn’
  - Numbers: 엄마 ➔ 3571 / 숫자 연쇄 (3571, 26, …)

도 입

숫자로 정의된 의미

- 엄마: 3571 *임의의 숫자/빈도수/…
  - 해당 표현의 의미를 얼마큼 잘 드러내는가?
  - 어휘간 의미적 연관성을 얼마큼 잘 포착하는가?
  - ‘의미계산’이 얼마큼 가능해지는가?
  - ….
- *mother*: obj/*tell* obj/*die* mod/*lone*
  - (27.49, 40.23, 59.05)
- 엄마
  - (-0.959987, 1.875226, -0.835720, 0.472719,

Sketch Engine

mother **British National Corpus freq = 26965**

object_of	3802	1.3
tell	204	27.49
marry	38	24.29
visit	57	24.29
ask	120	23.36
say	310	22.52
remember	59	22.45
help	78	21.4
kill	49	21.07
see	194	20.68
murder	18	19.83
kiss	17	19.69
ring	24	18.97
phone	15	18.47
nurse	10	17.94
hear	49	17.21

subject_of	5552	3.8
die	247	40.23
say	476	24.64
tell	159	21.61
live	76	21.48
breast-feed	6	21.04
cry	27	20.11
come	164	19.62
complain	21	18.33
speak	43	17.89
go	156	16.84
look	95	16.68
marry	22	16.61
weep	9	16.17
love	27	15.75
sit	40	15.67

adj_subject_of	680	2.5
ill	31	33.6
dead	26	27.3
alive	16	24.93
upset	9	21.96
likely	23	19.61
fond	6	18.1
married	9	17.89
worried	8	17.86
able	20	17.52
kind	5	16.63
right	15	15.79
happy	10	15.76
shocked	5	15.61
pleased	7	15.54
busy	7	14.96

modifier	3463	0.4
lone	163	59.05
queen	268	52.46
widowed	63	50.59
foster	83	49.38
unmarried	69	48.1
expectant	37	44.18
surrogate	36	41.31
teenage	58	39.83
single	153	35.85
working	80	34.01
young	158	32.89
distraught	12	25.56
poor	55	24.68
dear	27	24.49
primal	12	24.37

워드임베딩

벡터로 표상된 의미: word embedding

- (low) dimensional (200 features/dimensions/ranks)

2	엄마/NNG	(-0.959987, 1.875226, -0.835720, 0.472719, -0.905178, 0.588503, -1.070872, -1.3
3	아빠/NNG	(-1.221776, 0.818246, 0.069205, 0.862084, -0.589856, -1.342358, -1.065546, -2.2
4	아들/NNG	(-0.513450, 2.360061, -0.670642, -4.023421, 1.661846, 1.367789, -0.965055, -2.6
5	딸__01/NNG	(0.228678, 1.906885, -1.636114, -3.534212, 1.572831, 0.719615, -0.457926, -2.30
6		
7	별로__01/MAG	(0.518515, -0.058747, 0.092773, -2.011054, -0.206037, 0.153440, -0.450727, -1.7
8	전혀__01/MAG	(0.380255, 1.446430, -0.383245, 0.527952, -0.520785, -0.338759, -2.268026, -1.3
9	아무것/NNG	(-0.016211, -0.310664, -0.139886, 0.348985, -0.515651, 1.058252, 0.349367, -2.1
10	아무런/MM	(-0.445376, 1.806085, -1.331323, -1.523020, 0.123285, 0.014299, -0.761506, -2.0
11		
12	이/JKS	(0.787377, 0.617836, -1.221694, -0.480646, -1.447715, -1.278593, -1.008080, 0.9
13	가/JKS	(-0.325049, -0.996780, -0.004611, -1.288200, 0.269098, 0.106555, 0.013025, -1.0
14	은/JX	(0.175985, 0.173772, -1.736717, -0.485608, -0.653281, -0.847429, -0.973180, 0.1
15	는/JX	(-1.120614, -0.892748, -1.011558, -1.889204, 0.347976, -0.070462, -0.266430, -1
16	는/ETM	(1.286125, -1.061289, -0.193473, 0.509479, -0.625730, -1.329989, 0.850474, 0.30

Word2vec: Word embedding tool

- Word2Vec is a group of related models that are used to produce word embeddings. (From Wikipedia)
  - These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.
  - Input: a large corpus of text
  - Output: a vector space (typically of several hundred dimensions), with each unique word in the corpus being assigned a corresponding vector in the space.
  - Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space
- 절차: 텍스트 전처리, 프로그램 설치 및 구동, 결과검색

Vector 계산 (in R)

- > head(x)
- [1] 0.518515 -0.058747 0.092773 -2.011054 -0.206037 0.153440
- > head(y)
- [1] 0.380255 1.446430 -0.383245 0.527952 -0.520785 -0.338759

Matrix

- > head(z)
- [,1] [,2]
- [1,] 0.518515 0.380255
- [2,] -0.058747 1.446430
- [3,] 0.092773 -0.383245
- [4,] -2.011054 0.527952
- [5,] -0.206037 -0.520785
- [6,] 0.153440 -0.338759

Similarity (유사도)

- Similarity between vectors: cosine similarity

- > cosine(z)
-         [,1]   [,2]
- [1,]   1.0000000   0.6913215
- [2,]   0.6913215   1.0000000

결과 자료

Cosine similarity

```
> cosine(z1)
              별로_01/MAG 전혀__01/MAG 아무것/NNG 아무런/MM
별로_01/MAG     1.0000000    0.6913215  0.4695367 0.4804113
전혀__01/MAG    0.6913215    1.0000000  0.4858505 0.5953088
아무것/NNG      0.4695367    0.4858505  1.0000000 0.4624520
아무런/MM       0.4804113    0.5953088  0.4624520 1.0000000
```

↑

별로__01/MAG	(0.518515, -0.058747, 0.092773, -2.011054, -0.206037, 0.153440, -0.450727, -1.7
전혀__01/MAG	(0.380255, 1.446430, -0.383245, 0.527952, -0.520785, -0.338759, -2.268026, -1.3
아무것/NNG	(-0.016211, -0.310664, -0.139886, 0.348985, -0.515651, 1.058252, 0.349367, -2.1
아무런/MM	(-0.445376, 1.806085, -1.331323, -1.523020, 0.123285, 0.014299, -0.761506, -2.0

결과 자료

Cosine similarity

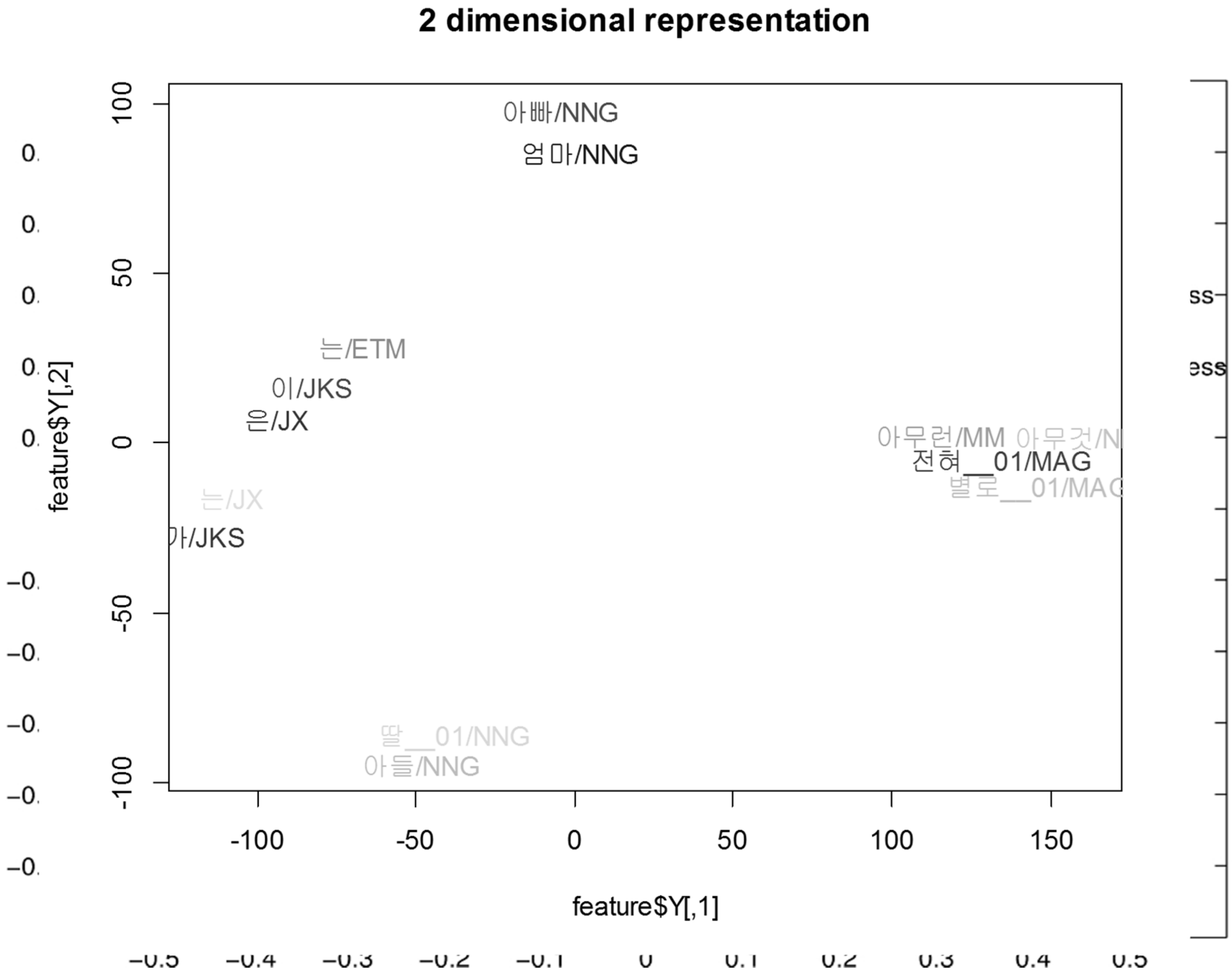
```
> cosine(z2)
             엄마/NNG  아빠/NNG  아들/NNG 딸__01/NNG
엄마/NNG    1.0000000 0.7996977 0.4203425  0.5058436
아빠/NNG    0.7996977 1.0000000 0.3108183  0.3855048
아들/NNG    0.4203425 0.3108183 1.0000000  0.8428138
딸__01/NNG 0.5058436 0.3855048 0.8428138  1.0000000
```

○

```
> cosine(z3)
            이/JKS    가/JKS     은/JX     는/JX    는/ETM
이/JKS 1.0000000 0.3687333 0.7221617 0.2666441 0.4636197
가/JKS 0.3687333 1.0000000 0.1531184 0.6493893 0.3426413
은/JX  0.7221617 0.1531184 1.0000000 0.5045738 0.3482699
는/JX  0.2666441 0.6493893 0.5045738 1.0000000 0.2578213
는/ETM 0.4636197 0.3426413 0.3482699 0.2578213 1.0000000
```

○

결과 자료


2 dimensional representation
아빠/NNG
엄마/NNG
는/ETM
이/JKS
은/JX
는/JX
가/JKS
아무런/MM
아무것/N
전혀__01/MAG
별로__01/MAG
딸__01/NNG
아들/NNG
feature$Y[,1]
feature$Y[,2]
-100
-50
0
50
100
150

결과 자료

Similarity: '별로' (역순 정렬)

- Word: 별로__01/MAG Position in vocabulary: 2661

Word	Cosine distance
전혀__01/MAG	0.759114
그다지/MAG	0.721255
별__02/MM	0.716306
별다르__01/VA	0.596534
딱히__02/MAG	0.594654
도/JX	0.594131
그리__02/MAG	0.592236
아무것/NNG	0.582232
별반__01/MAG	0.574758
특별히/MAG	0.573280
거의__01/MAG	0.568257
아무__01/MM	0.556437
아무런/MM	0.540597
아무래도/MAG	0.530550
꽤__01/MAG	0.516622
도무지__02/MAG	0.506698
밖에/JX	0.505642
아무__01/NP	0.500108
별다르/VA	0.495149
좀처럼/MAG	0.490569
적__02/VA	0.489714
못하/VA	0.484308
절대__05/MAG	0.483904
썩__01/MAG	0.475421
아직__01/MAG	0.474474

결과 자료

Word embedding 활용 예

- Synonymous/antonymous word list for a given word
- Semantic classes
  - V4908 8 거의__01/MAG 전혀__01/MAG 별로__01/MAG 이루__01/MAG 별반__01/MAG 딱히__02/MAG 도저히/MAG 별달리/MAG
- Inference/analogy
  - <u>한국 : 서울 = [   ] : 도쿄</u> *한국-서울+도쿄=
  - <u>아빠 : 아들 = [   ] : 딸</u>

결과 자료

단일어휘 결합 기능

- maybe 900
- i_guess:0.81 really:0.78 little_bit:0.78 probably:0.76 i_suppose:0.75 just:0.75 feel_like:0.74 definitely:0.74 something:0.73 that's:0.73 it's:0.73 you:0.73 anyway:0.72 something_else:0.72 think:0.72 you_know:0.72 i'd:0.71 i_think:0.71 so:0.71 you'd:0.71 i:0.71 sure:0.71 bit:0.71 i'm_sure:0.70 thing:0.70 things:0.70 i_don't_know:0.70 going:0.69 we'd:0.69 okay:0.69 ok:0.69 lot:0.69 it'll:0.69 nice:0.69 i'll:0.69 yeah:0.68 know:0.68 me:0.68 you've_got:0.68 wonder_if:0.67

결과 자료

- Word: 가공유NNG Position in vocabulary: 145420
- Word   Cosine distance
- --
- 기능성NNG_우유02NNG 0.599173
- 밀크NNG_플러스NNG 0.594060
- 우유02NNG 0.586016
- 유제품NNG 0.579932
- 칼슘NNG_우유02NNG 0.578621
- 매일유업NNP 0.567865
- 가공01NNG_우유02NNG 0.564718
- 프렌치NNG_카페NNG_카페NNG_믹스NNG 0.563719
- 초코NNP_우유02NNG 0.558722
- 커피NNG_믹스NNG 0.553025
- 산양01NNG_분유04NNG 0.549890
- 유86NNG_가공01NNG_업체NNG 0.545099
- 흰우유NNG 0.543447
- 모유01NNG_성분01NNG 0.535308
- 유산균NNG_발효유NNG 0.534920

결과 자료

- Word: 카페오레NNG Position in vocabulary: 325481

Word	Cosine distance
라테NNG	0.631076
프렌치NNG_카페NNG	0.616583
아메리카02NNP_놀01VV	0.614844
커피NNG	0.614594
아메리카02NNP_노12NNG	0.609900
카페NNG_모카NNG	0.602005
카푸치노NNG	0.600762
라떼NNP	0.600240
요거01NP_트01VV	0.596785
요거트NNG	0.595597
녹차01NNG_라떼NNP	0.592509
카라멜NNG	0.588655
고구마NNG_케이크NNG	0.585035
네스99NNP_카페NNG	0.581325
아이스NNG_아메리카02NNP_노11NNP	0.579934

결과 자료

- Word: 무사03NNG_바03NNB_예JX Position in vocabulary: 474509

Word	Cosine distance
소유03NNG_즈03NP	0.563065
소유즈호NNG	0.533329
러시아NNP_우주02NNG_비행사NNG	0.516233
러시아NNP_항공NNG_우주국NNG	0.498318
바이코누르NNP_우주02NNG_기지08NNG	0.485914
우주02NNG_왕복NNG_선19XSN_소유즈호NNG	0.473509
국제02NNG_우주02NNG_정거장NNG	0.470252
러시아NNP_우주선01NNG	0.466425
말렌첸코NNG	0.464894

언어학적 활용?

문장 의미는?

- 엄마가 오셨다.
  - 엄마 (-0.959987, 1.875226, -0.835720, 0.472719,
  - 가 (-0.325049, -0.996780, -0.004611, -1.288200,
  - 오 (-0.435634, -0.660927, 1.567977, -2.256657,
  - 시 (-0.129665, 0.907324, 1.499305, -2.020838,
  - 었 (1.125780, 1.085099, -0.499714, -0.365474,
  - 다 (2.053955, 0.144086, -0.757581, 0.490556,
- Compositionality?

언어학적 활용?

Distributional semantics

- "Distributional semantics is a theory of meaning which is computationally implementable and very, very good at modelling what humans do when they make similarity judgements. … This approach to meaning is in no way the only one, but has come from a particular philosophical tradition involving linguists and philosophers such as **Leonard Bloomfield**, **Zellig Harris**, **J.R. Firth** or again **Ludwig Wittgenstein** (in his later work) and **Margaret Masterman**.
  - http://aurelieherbelot.net/research/distributional-semantics-intro/

언어학적 활용?

Compositional distributional semantics!

- "Compositional distributional semantic models are an extension of distributional semantic models that characterize the semantics of entire phrases or sentences. This is achieved by composing the distributional representations of the words that sentences contain. Different approaches to composition have been explored, and are under discussion at established workshops such as SemEval. From Wikipedia

언어학적 활용?

언어학에서의 활용방안

- 큰 질문

- Downloadable pre-trained word vectors
  - https://nlp.stanford.edu/projects/histwords/
  - https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

참고자료

- 이용 자료
  - 세종 의미분석 말뭉치; 물결21(http://corpus.korea.ac.kr) 일부
- 이용도구
  - word2vec (Ubuntu Linux환경), R, Perl
- 참고문헌

  - Wikipedia
  - François Chollet. 2018. *Deep Learning with R*. Manning.
  - Tim Berners Lee. 1999. *Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*, Harper San Francisco.
  - Tim Berners-Lee. 2001 . "The Semantic Web". *Scientific American*: May 17, 2001.