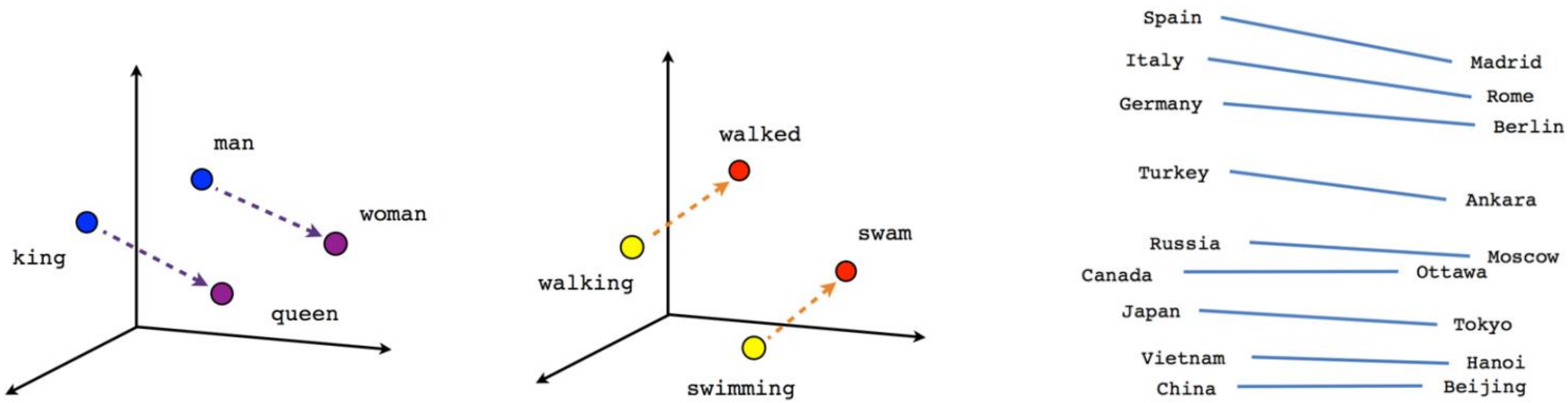


Male-Female Verb tense Country-Capital

Lecture 5: Text Representation II
Distributed Representations

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

GloVe

Pennington et al. (2014)

- Limitations of Word2Vec
  - ✓ The network spends so much time to train some overwhelmingly used words
    - Ex: to learn a distribution for P(w|the)

Theatre or theater is a collaborative form of fine art that uses live performers to present the experience of a real or imagined event before a live audience in a specific place. The performers may communicate this experience to the audience through combinations of gesture, speech, song, music, and dance. Elements of art and stagecraft are used to enhance the physicality, presence and immediacy of the experience. The specific place of the performance is also named by the word "theatre" as derived from the Ancient Greek (thatron, "a place for viewing"), itself from (theomai, "to see", "to watch", "to observe"). Modern Western theatre comes from large measure from ancient Greek drama, from which it borrows technical terminology, classification into genres, and many of its themes, stock characters, and plot elements. Theatre artist Patrice Pavis denes theatricality, theatrical language, stage writing, and the specificity of theatre as synonymous expressions that differentiate theatre from the other performing arts, literature, and the arts in general. Theatre today, broadly denfied, includes performances of plays and musicals, ballets, operas and various other forms.

GloVe

Pennington et al. (2014)

- GloVe
  - ✓ Based on matrix factorization method
  - ✓ http://nlp.stanford.edu/projects/glove/
  - ✓ Notations

- $X \in \mathbb{R}^{V \times V}$ word co-occurrence matrix
- X_{ij} frequency of word i co-occurring with word j
- $X_i = \sum_k^V X_{ik}$ total number of occurrences of word i in corpus
- $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ a.k.a. probability of word j occurring within the context of word i
- $w \in \mathbb{R}^d$ a word embedding of dimension d
- $\tilde{w} \in \mathbb{R}^d$ a context word embedding of dimension d

GloVe

Pennington et al. (2014)

- Motivation

Prob. and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$\frac{P(k\|ice)}{P(k\|steam)}$	8.9	8.5×10^{-2}	1.36	0.96

✓ For words *k* related ice but not steam (solid), the ratio P_{ik}/P_{jk} is large

✓ For words *k* related steam but not ice (gas) the ratio P_{ik}/P_{jk} is small

✓ For words *k* that are either related to both ice and steam, or to neither, the ratio should be close to 1

GloVe

Pennington et al. (2014)

- Formulation
  - ✓ Express the relationship among three words using a function F

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

  - ✓ Relationship between w_i and w_j is expressed by subtraction

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

  - ✓ Inner product is used to link $\tilde{w}_k$ with w_i and w_j

$$F\Big((w_i - w_j)^T \tilde{w}_k\Big) = \frac{P_{ik}}{P_{jk}}$$

GloVe

Pennington et al. (2014)

- Homomorphism

Prob. and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$\frac{P(k\|ice)}{P(k\|steam)}$	8.9	8.5×10^{-2}	1.36	0.96

✓ Want to preserve $\frac{P(k|ice)}{P(k|steam)}$ using $F\Big((w_i - w_j)^T \tilde{w}_k\Big) = \frac{P_{ik}}{P_{jk}}$

$$\frac{P(solid|ice)}{P(solid|steam)} = F\Big((ice - steam)^T solid\Big)$$

$$\frac{P(solid|steam)}{P(solid|ice)} = F\Big((steam - ice)^T solid\Big)$$

GloVe

Pennington et al. (2014)

- Homomorphism

$$F\Big((ice - steam)^T solid\Big) = \frac{P(solid|ice)}{P(solid|steam)} = \frac{1}{F\Big((steam - ice)^T solid\Big)}$$

$$(ice - steam)^T solid = -(steam - ice)^T solid$$

inverse element of addition

$$F\Big((ice - steam)^T solid\Big) = \frac{1}{F\Big((steam - ice)^T solid\Big)}$$

inverse element of multiplication

✓ Homomorphism preserves an operation, which in turn preserves the inverse element

✓ Need a homomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}_{>0}, \times)$

GloVe

Pennington et al. (2014)

- Homomorphism
  - ✓ Function F: homomorphism that maps $(\mathbb{R}, +)$ to $(\mathbb{R}_{>0}, \times)$

$$w_i^T \tilde{w}_k = (w_i - w_j)^T \tilde{w}_k + w_j^T \tilde{w}_k$$

$$\begin{aligned} F(w_i^T \tilde{w}_k) &= F\Big((w_i - w_j)^T \tilde{w}_k + w_j^T \tilde{w}_k\Big) \\ &= F\Big((w_i - w_j)^T \tilde{w}_k\Big) \times F(w_j^T \tilde{w}_k) \end{aligned}$$

$$F\Big((w_i - w_j)^T \tilde{w}_k\Big) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)} = \frac{P_{ik}}{P_{jk}}$$

  - ✓ Finally, we can drive that $F(x) = \exp(x)$

GloVe

Pennington et al. (2014)

- Solution
  - ✓ We know that $F(x) = \exp(x)$ and $F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)} = \frac{P_{ik}}{P_{jk}}$

$$\exp\left(w_i^T \tilde{w}_k - w_j^T \tilde{w}_k\right) = \frac{\exp\left(w_i^T \tilde{w}_k\right)}{\exp\left(w_j^T \tilde{w}_k\right)}$$

↓

$$w_i^T \tilde{w}_k = \log P_{ik} = \log X_{ik} - \log X_i$$

↓

$$w_i^T \tilde{w}_k = \log P_{ik} = \log X_{ik} - \log X_i$$

$$w_i^T \tilde{w}_k = \log X_{ik} - b_i - \tilde{b}_k$$

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log X_{ik}$$

GloVe

Pennington et al. (2014)

- Objective Function
  - ✓ A least squared objective function

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

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

where f has the following desiderata:

1. $f(0) = 0$
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.

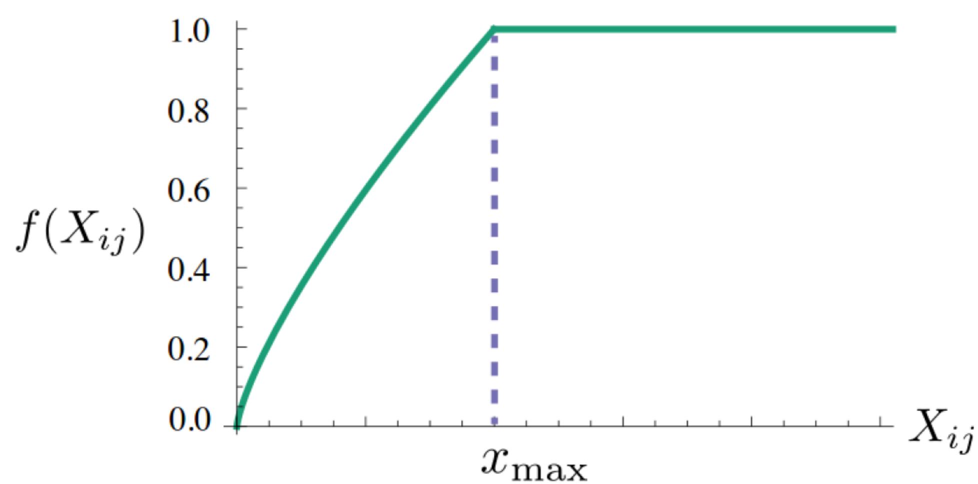
GloVe

Pennington et al. (2014)

- Objective Function
  - ✓ A least squared objective function

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

where $f(x) = \begin{cases} \left(\frac{x}{x_{\max}}\right)^{\alpha} & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$



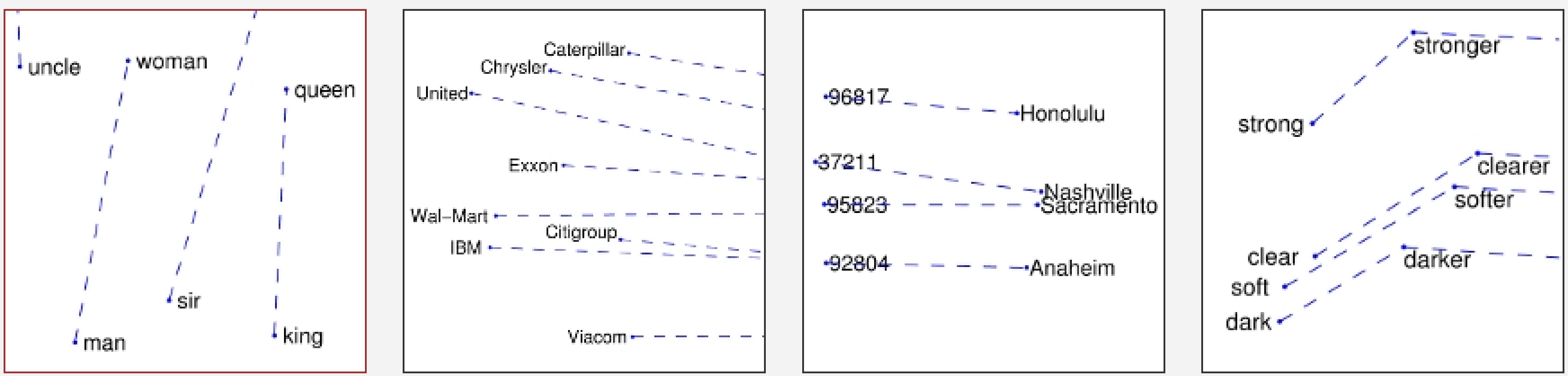
GloVe

Pennington et al. (2014)

- **Results**

0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

3. litoria 4. leptodactylidae 5. rana 7. eleutherodactylus



man - woman company - ceo city - zip code comparative - superlative

http://nlp.stanford.edu/projects/glove/

AGENDA

FastText

Bojanowski et al. (2016)

- Limitations of NNLM, Word2Vec, and GloVe
  - ✓ Ignores the morphology or words by assigning a distinct vector to each word
  - ✓ Difficult to apply to morphologically rich languages with large vocabularies and many rare words (Turkish or Finnish)
- Goal
  - ✓ Learn representations for character n-grams
  - ✓ Represent words as the sum of n-gram vectors

FastText

Bojanowski et al. (2016)

- Revisit Negative Sampling in Word2Vec

$$J_t(\theta) = \log\sigma(u_o^T v_c) + \sum_{i=1}^{k} E_{i\sim P(w)}\left[\log\sigma(-u_i^T v_c)\right]$$

✓ Score is just a dot product between the two embeddings

- Subword model

✓ Define the set of n-grams appearing in w: $\mathcal{G}_w \subset \{1, ..., G\}$

$$score(w, c) = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g^T \mathbf{v}_c$$

✓ Represent a word by the sum of the vector representations of its n-grams

FastText

Bojanowski et al. (2016)

- Subword model
  - ✓ n-gram representation
    - Include the word w in the set of its n-grams
    - Keep all the n-grams of size 3, 4, 5, and 6
    - Different vectors are assigned to a word and a n-gram sharing the same sequence of characters

Word2Vec

parameter	...	...	...	...	...

FastText

p	...	...	...	...	...
...	...	...	...	...	...
r	...	...	...	...	...
pa	...	...	...	...	...
...	...	...	...	...	...
er	...	...	...	...	...
par	...	...	...	...	...
...	...	...	...	...	...
ameter	...	...	...	...	...
Avg.	...	...	...	...	...

Word Embedding Examples

- Word Embedding examples: English

✓ Word lists that are close to a given word after embedding

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

Collobert et al. (2011)

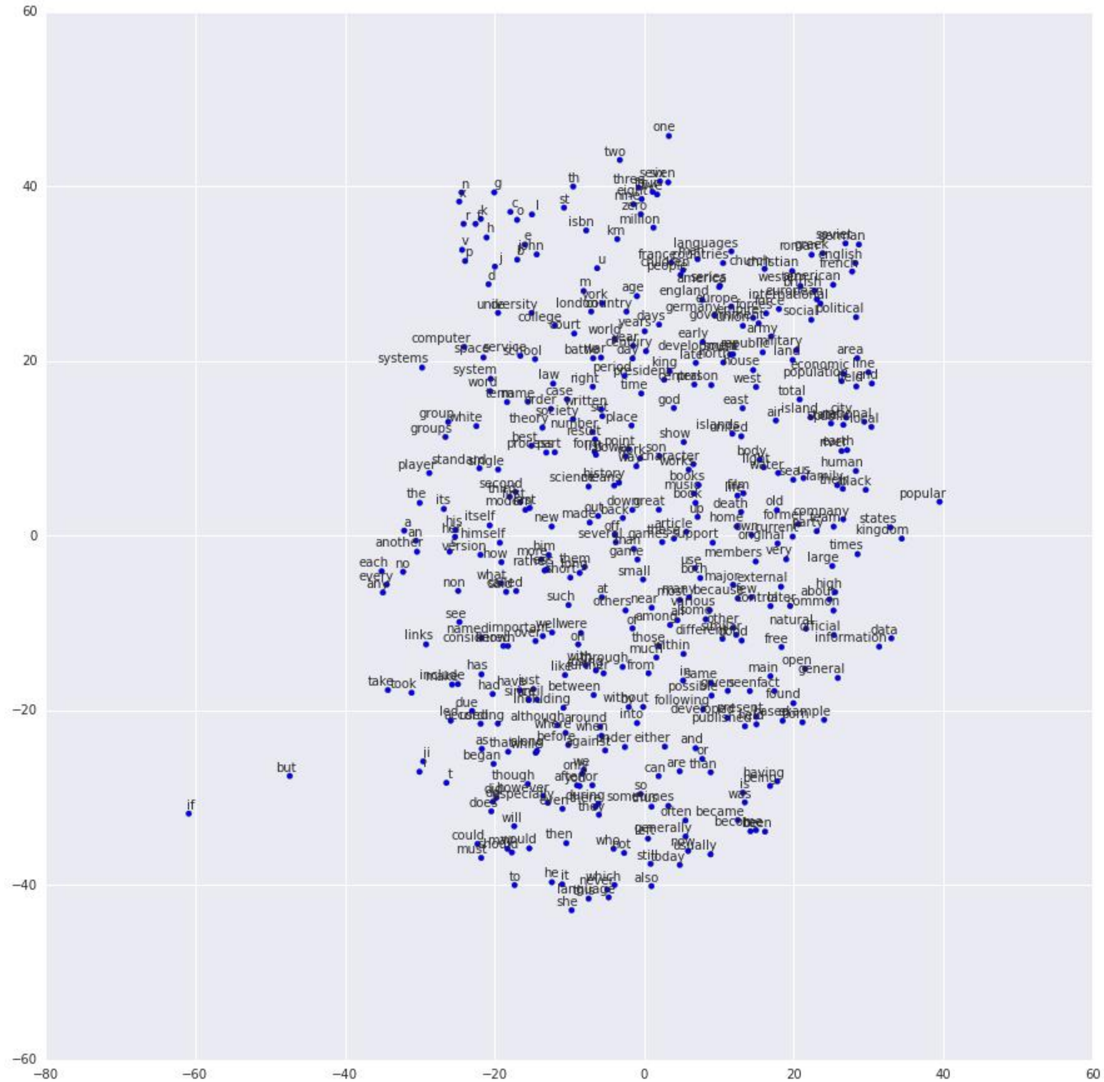
✓ Relationship pairs in a word embedding

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Mikolov et al. (2013)

Word Embedding Examples

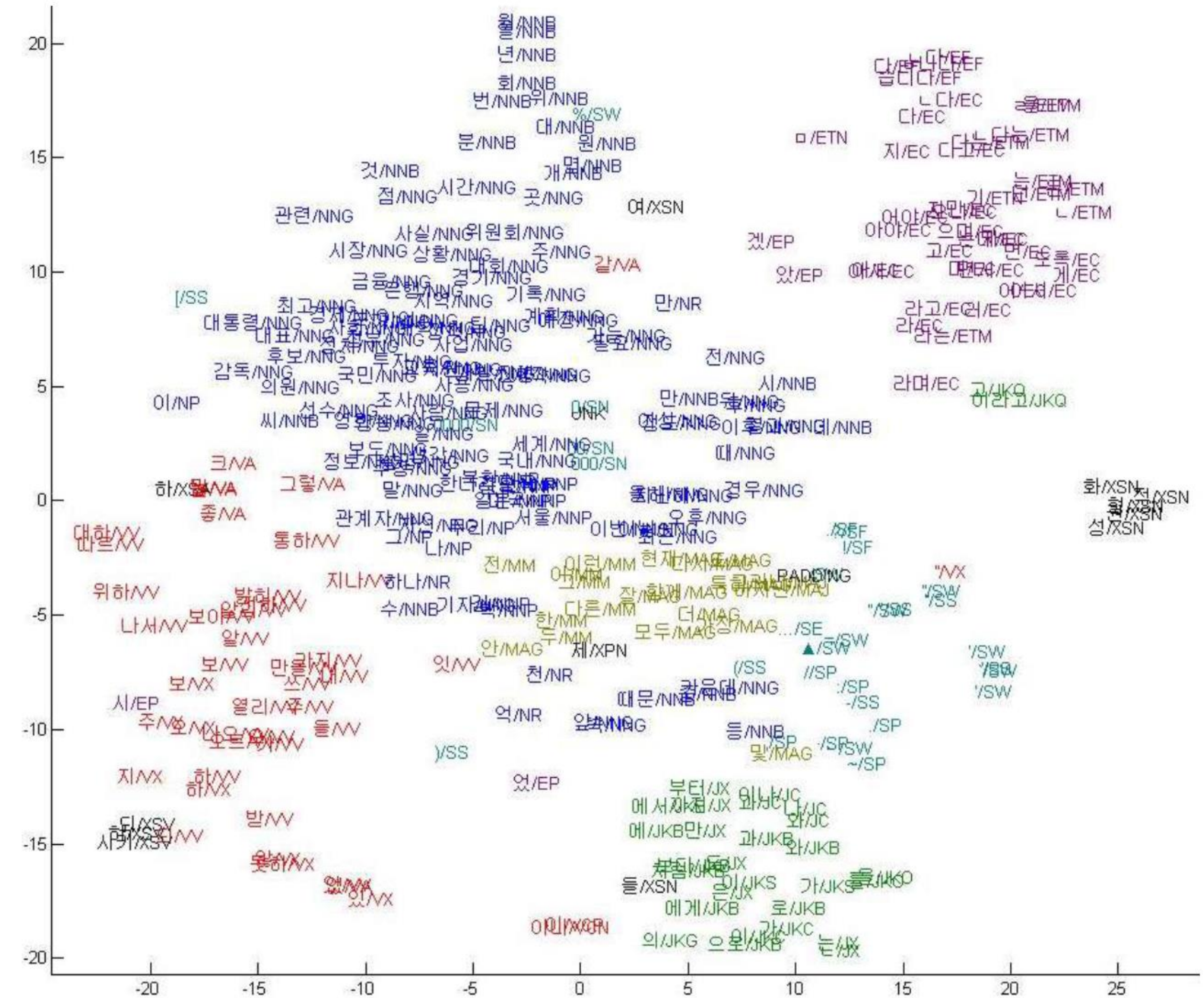
- Word Embedding examples: English



https://www.tensorflow.org/versions/r0.7/tutorials/word2vec/index.html

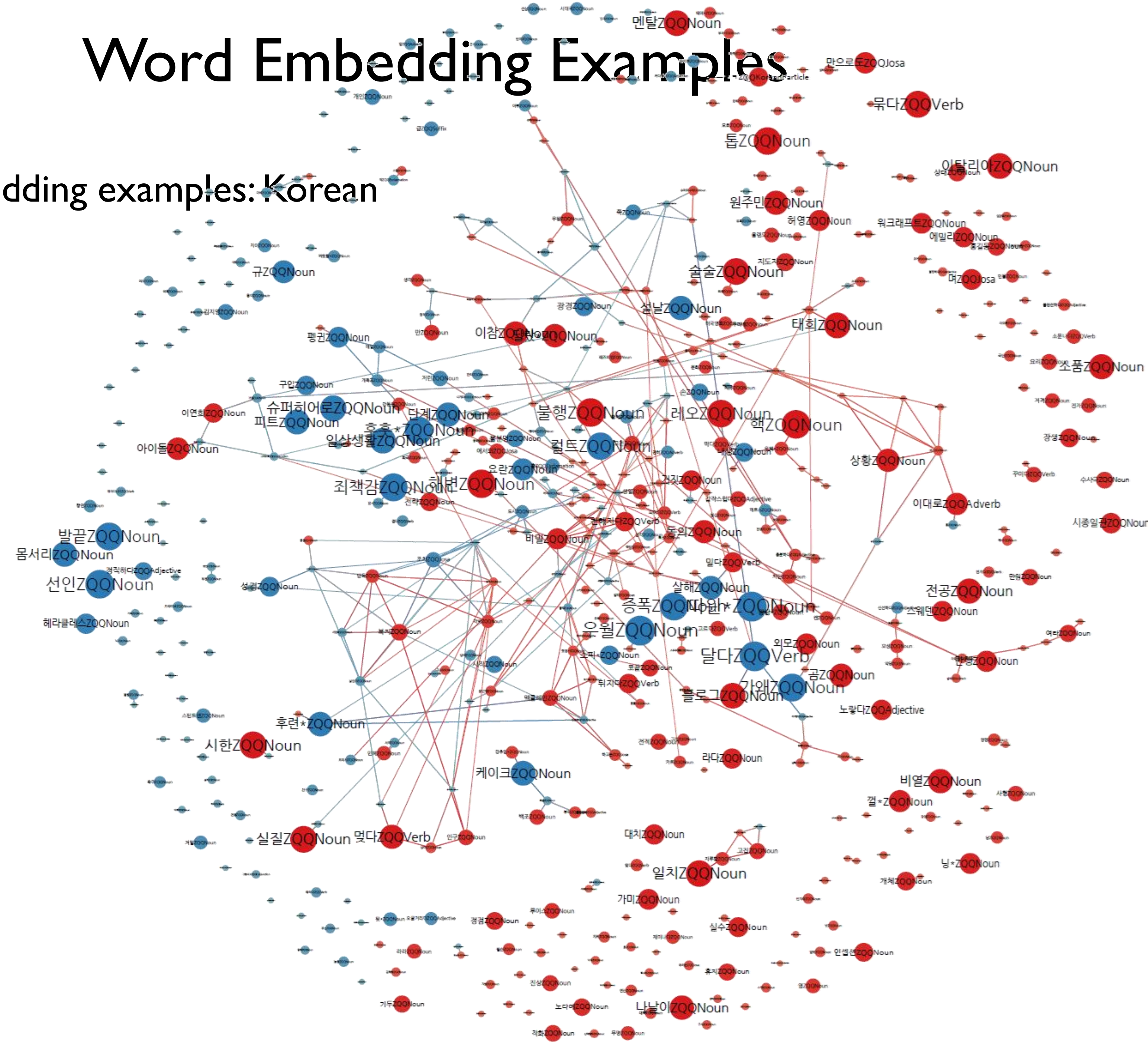
Word Embedding Examples

- Word Embedding examples: Korean



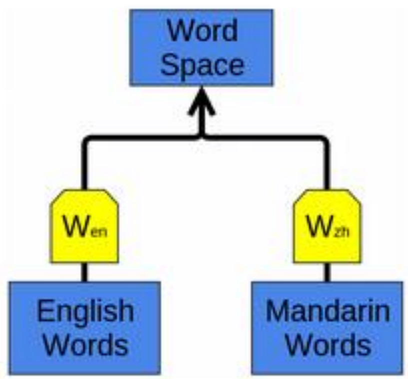
Word Embedding Examples

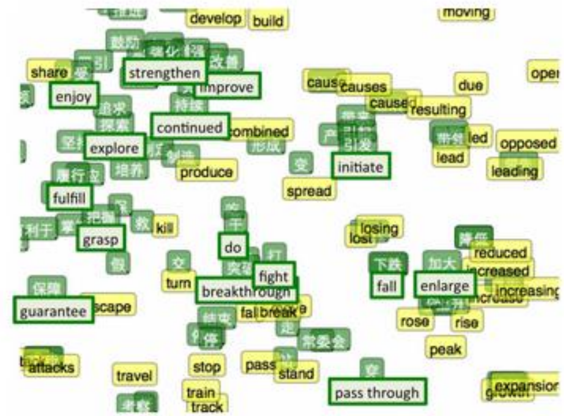
- Word Embedding examples: Korean



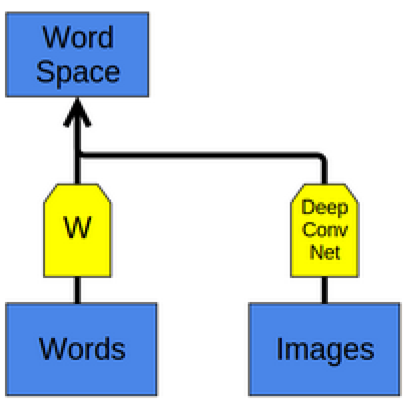
Word Embedding Examples

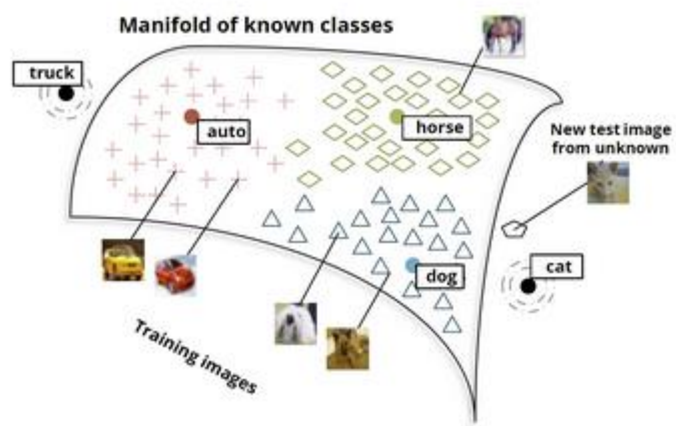
- Word Embedding with two different languages

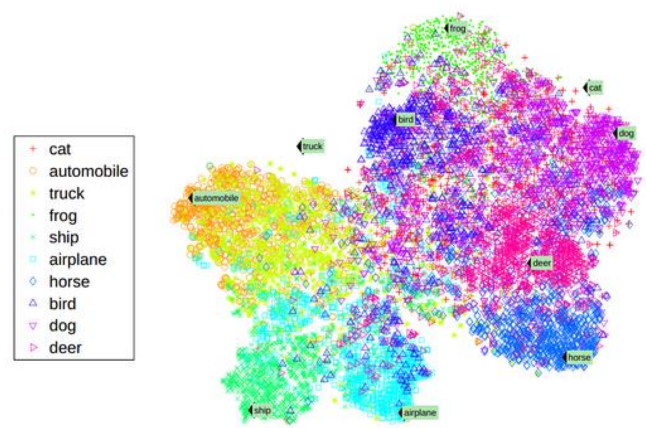




- Word Embedding with Images








ANY
questions?