COMP90042 Web Search & Text Analysis

Workshop Week 5

Zenan Zhai

April 9, 2019

University of Melbourne

Warm up

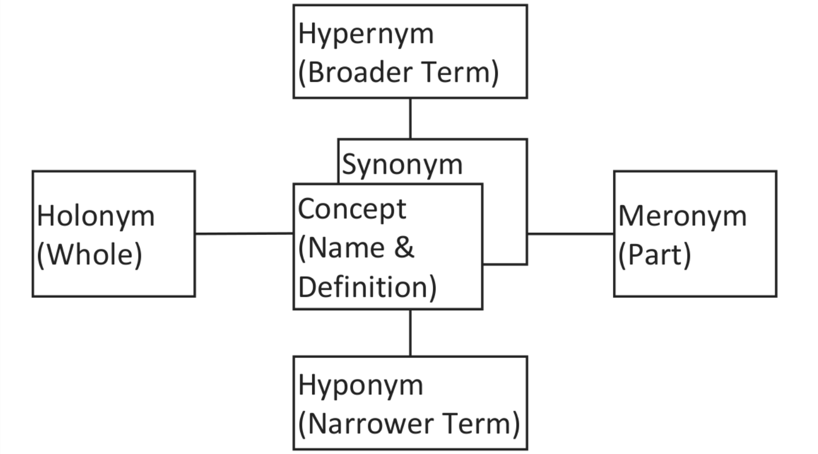
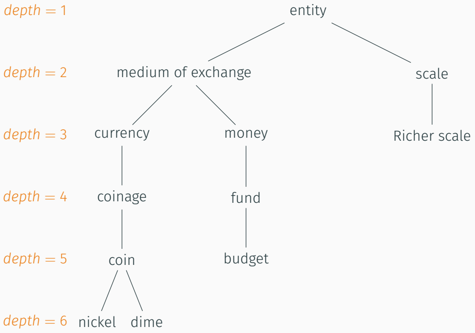


Fig from workshop slides 2018 by Yuan Li
`http://vlearn.fed.cuhk.edu.hk/meaningrelations/srex_meronyms/`

Outline

- Lexical Semantics
    - Wu & Palmer Similarity
    - Lin Similarity
    - Word Sense Disambiguation (WSD)
- Distributional Semantics
    - Vector Space Model (VSM)
    - Point-wise Mutual Information (PMI)
    - Neural Word Embeddings
- Excercise - Notebook

Wu & Palmer Similarity



$$Simwup = \frac{2 \times depth(LCS(w_1, w_2))}{depth(w_1) + depth(w2)}$$

- Lowest Common Subsummer (LCS)
  - Deepest shared parent of 2 words.
- Use depth instead of path length.
- Exercise *information* and *retrieval*

Lin Similarity

Information Content

$$IC(w) = -logP(w)$$

- rare words are more informative

Lin Similarity

$$Simlin(w_1, w_2) = \frac{2 \times IC(LCS(w_1, w_2))}{IC(w_1) + IC(w2)}$$

- penalize frequent LCS
- penalize similarity between rare words

Word Sense Disambiguation

Definition

- Automatically determining which sense (usually, Wordnet synset) of a word is intended fora given token instance with a document.

Supervised Methods

- Trained classifier for choosing the correct sense

Less supervised Methods

- Lesk - match WordNet dict gloss to context
- Yarowsky - Bootstrap adding confident prediction to training set

Outline

- Lexical Semantics
    - Wu & Palmer Similarity
    - Lin Similarity
    - Word Sense Disambiguation (WSD)
- Distributional Semantics
    - Singular Vector Decomposition (SVD)
    - Point-wise Mutual Information (PMI)
    - Neural Word Embeddings
- Excercise - Notebook

PMI - Concept

$$PMI(a, b) = log_2 \frac{P(a, b)}{P(a) \times P(b)}$$

Numerator: Actual joint prob. observed in corpus.

- More weight for words appears in pairs.

Denominator: Expect prob. under independent assumption.

- Penalize frequent words (e.g. the, a)

Weakness

- When $P(a, b) = 0, PMI(a, b) = -\infty$
- Bias for co-occurrence of 2 rare words.

PMI - Excercise

	apple	pear	banana	peach	$\sum$
fruit	3	0	4	1	8
delicious	0	3	0	0	3
bad	0	0	4	4	8
company	1	2	0	0	3
$\sum$	4	5	8	5	22

$$P(fruit) = \frac{8}{22}$$

$$P(apple) = \frac{4}{22}$$

$$P(fruit, apple) = \frac{3}{22}$$

$$PMI(fruit, apple) = \frac{P(fruit, apple)}{P(fruit) \times P(apple)}$$

PMI - Exercise

	cup	(not) cup
world	55	225
(not) world	315	1405

- PMI(world, cup) ?
- How to get word vectors from PMI?

Neural Word Embeddings

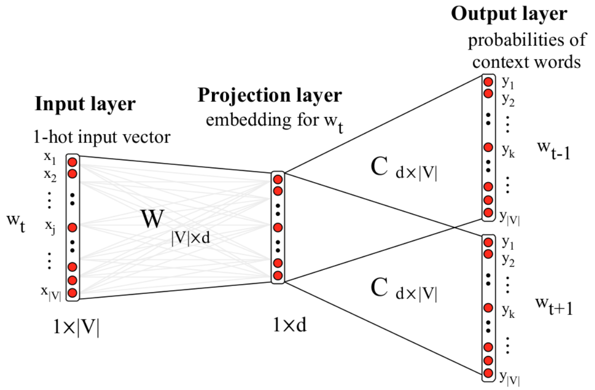


Fig 19.18, JM

- What is Softmax?
- What is Negative Sampling?
- Where does the word embeddings come from?
- What is the difference between CBOW and skip-gram?