

NATURAL LANGUAGE PROCESSING

LESSON 9 : SEMANTIC SIMILARITY

OUTLINE

- Semantic Relations
- Semantic Similarity Levels
 - Sense Level
 - Word Level
 - Text Level
- WordNet-based Similarity Methods
- Hybrid Methods Similarity
- PageRank-based Similarity

SEMANTIC RELATIONS

- In the literature, three different terms are used for similarity: semantic relatedness, similarity and semantic distance.
- There are many semantic relation types.
- The most important semantic relation is **synonym** (black-dark) and **antonym** (black-white).
- But non-synonym (or non-antonym) entities may also be semantically related by other relationships such as **meronym** (finger is meronym of hand), **hyponym** (eagle is hyponym of bird), **hypernym** (bird is hypernym of eagle).

SEMANTIC RELATIONS

Relation type	Example
Synonym (synset)	Different - Unlike
Antonym	Buy – Sell
Category Domain	Cell - Biology
Sub event	Search – Query
Causes	Slimming, Weight loss
Hypernymy	Jam – Rose Jam
Hyponymy	Rose Jam – Jam
Similar to	Next – Following

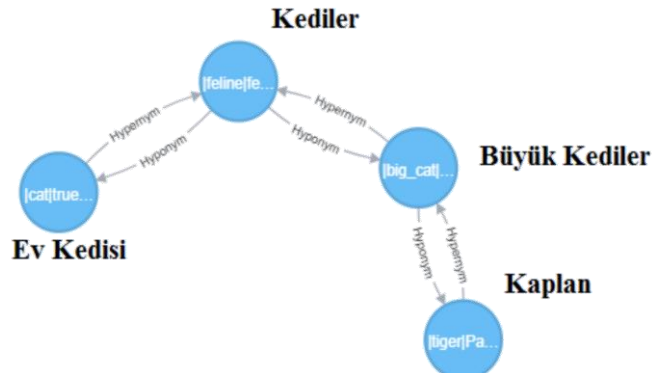
SEMANTIC RELATIONS

- Semantic similarity relates to computing the conceptual similarity between terms according to a lexicon (dictionary).

- Example:

«tiger» «cat»

«kaplan» «ev kedisi»



SEMANTIC RELATIONS

- While *Antarctica* and *penguin* are not similar according to their lexical definitions, we feel a strong relation between them.
- Because most of penguins **live in** Antarctica. But there is not **'live in'** relation among the known semantic relations. For this kind of relations we should define a general relation such as **'related'**.
- Note that semantic **relatedness** is a more flexible relation than the other known ones.
- For some other word pairs, **related** can be used (pencil–paper, penguin–Antarctica, rain–flood).

SEMANTIC SIMILARITY LEVELS


There are three type of semantic similarity levels:

- **Sense level** deals with the conceptual part of a word. It is a unique representation of a concept and has no ambiguity.
- **Word level** deals with the word which might contain multiple senses, so ambiguity can be possible.
- **Text level** including short text (sentence, paragraph) and documents. In this level, a text has usually several ambiguity.

SENSE LEVEL SIMILARITY


- It is the primary step of similarity, sense is the concept that a word aims to define.
- A typical sense fox#n#1, n (noun) is part of speech tagging and 1 is the first meaning in dictionary.
 - fox#n#1: alert carnivorous mammal.
 - fox#n#2: a shifty deceptive person.
- To understand a text in sense level, at first, it requires word sense disambiguation.

SENSE LEVEL SIMILARITY

- Sense-level semantic similarity are mostly based on lexical resources such as dictionary or thesaurus.
 - Lexical resources are mostly used in form of semantic networks.
 - In order to determine semantic similarity of two words, it is used structural properties of these kind of networks such as adjacencies.
 - The most popular lexical resource is the **WordNet**.
- 


SENSE LEVEL SIMILARITY

In addition to WordNet, other lexical resources:

- Collaboratively-constructed resources such as
 - Wikipedia
 - Wiktionary
 - Dictionaries such as
 - Longman Dictionary
 - Integrated knowledge resources such as
 - BabelNet
- 


WORD LEVEL SIMILARITY

The approaches at the word level can be grouped into two categories:

- Distributional approaches
 - Lexical resource-based approaches
- 

WORD LEVEL SIMILARITY

Distributional approaches use co-occurrence statistics for the computation of vector-based representations of different words.

- The weights in co-occurrence-based vectors are usually computed by means of TF-IDF or Pointwise Mutual Information (PMI).
 - The dimensionality of the resulting weights matrix is often reduced, for instance using Singular Value Decomposition.
 - The structured textual content of specific lexical resources such as Wikipedia has been used with distributional approaches.
- 

WORD LEVEL SIMILARITY


Lexical resource-based approaches usually assume that;

- The similarity of two words can be calculated in terms of the similarity of their closest senses.

For example, fox and rabbit


dodger, fox, slyboots: a shifty deceptive person

fox: the grey or reddish-brown fur of a fox

- Therefore every sense-level approach can be directly applied for word similarity.
- 

TEXT LEVEL SIMILARITY

Text-level similarity methods can be grouped into two categories:

- Viewing a text as a combination of words and calculate the similarity of two texts by aggregating the similarities of word pairs across the two texts,
 - Modelling a text as a whole and calculate the similarity of two texts by comparing the two models obtained.
- 

TEXT LEVEL SIMILARITY

Approaches in the first category search for pair of words in different texts that maximize similarity and compute the overall similarity by aggregating individual similarity values.

- *‘Car goes faster than horse.’* tokens={car, go, fast, horse}
- *‘Train goes in railway.’* tokens={train, go, railway}


$$\text{Similarity}(S_1, S_2) = \frac{\sum_1^n \sum_1^m \max(\text{sim}(T_n, T_m))}{n}$$

TEXT LEVEL SIMILARITY


The second category usually involves transforming texts into vectors and computing the similarity of texts by comparing their corresponding vectors.

- Vector space models are an example of this category.
- Models mainly focused on the representation of larger pieces of text, such as documents, where a text is modeled on the basis of the frequency statistics of the words it contains.
- Such models, however, suffer from sparseness and cannot capture similarities between short text pairs that use different wordings.

WORDNET BASED SIMILARITY METHODS

- WordNet is the most common structural lexical knowledge and organized hierarchically in graph structure.
 - It consist of nodes and edges, nodes are the **synsets** and edges are the **relations**.
 - More than 20 relation types are defined in the current version of the WordNet.
 - WordNet based first methods use Hypernym (Is a), Meronymy (Part of) and Antonymy relations.
 - Recent methods uses all available relation types.
- 

WORDNET BASED SIMILARITY METHODS

- WordNet based similarity methods use graph structure of the WordNet, and measures similarity using several metrics like path length, depth length, lcs (lowest common subsumer), direction of the relations.
 - Following methods are the most frequently used WordNet Based similarity methods.
 - Leacock & Chodorow Method (1998)
 - Wu & Palmer Method – (1994)
 - Hirst And St-onge Method (1998)
- 

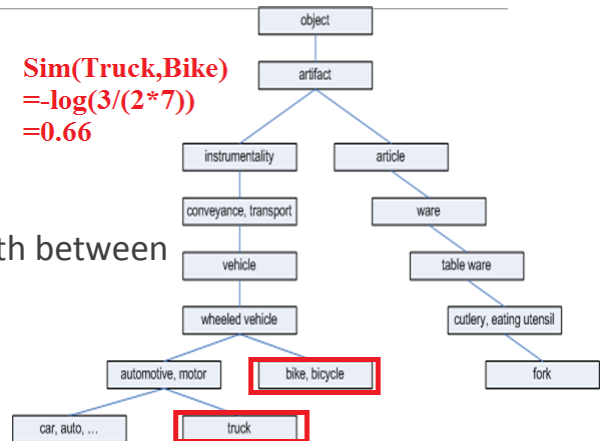
LEACOCK & CHODOROW METHOD (1998)

$$SIM_{LC}(A, B) = -\log \frac{Len(A, B)}{2 * D_{max}}$$

Len(A,B) : length of the shortest path between two concepts using node-counting

D_{max} : max depth of the taxonomy

$$\begin{aligned} Sim(Truck, Bike) &= -\log(3/(2*7)) \\ &= 0.66 \end{aligned}$$



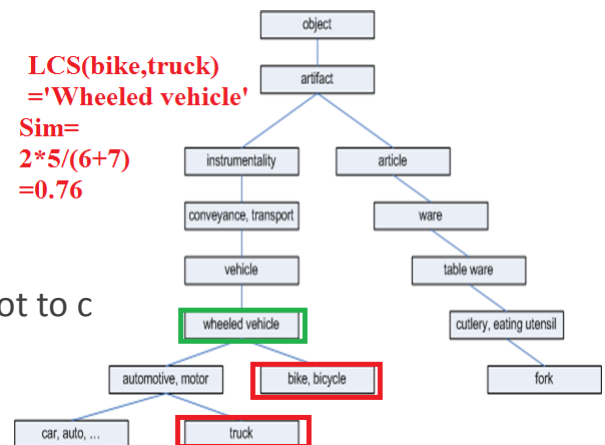
WU & PALMER METHOD – (1994)

$$Sim_{W\&P}(c1, c2) = \frac{2 * depth(lcs(c1, c2))}{depth(c1) + depth(c2)}$$

lcs(a, b): lowest common subsumer

depth(c) : number of edges from root to c

$$\begin{aligned} LCS(bike, truck) &= 'Wheeled vehicle' \\ Sim &= \frac{2 * 5}{6 + 7} \\ &= 0.76 \end{aligned}$$



HIRST AND ST-ONGE METHOD (1998)

Hirst and St-Onge's approach is summarized by the following formula for two WordNet concepts $c1 \neq c2$:

$$\text{relHS}(c1, c2) = C - \text{len}(c1, c2) - k \times \text{turns}(c1, c2)$$

where C and k are constants (in practice, they used $C = 8$ and $k = 1$), $\text{turns}(c1, c2)$ is the number of times the path between $c1$ and $c2$ changes direction.

$$\text{relHS}(\text{bike}, \text{truck}) = 8 - \text{len}(\text{bike}, \text{truck}) - \text{change_of_direction}$$

$$\text{relHS}(\text{bike}, \text{truck}) = 8 - 3 - 1 = 4$$

Here, the maximum similarity is 8 and the minimum is 0.

HYBRID METHODS FOR SIMILARITY

Hybrid methods utilize both WordNet and a corpus. Also known as information content based methods.

- It gathers information from corpus for a specific concept, and uses WordNet to get similarity.
- Information Content (IC) is a measure of specificity for a concept.
- Higher values are associated with more specific concepts (e.g., pitchfork: yaba), while those with lower values are more general (e.g., idea: fikir).
- IC is computed based on frequency counts of a concepts found in a text corpus.

HYBRID METHODS FOR SIMILARITY

- For each concept (synset) c in WordNet, Information Content is defined as the negative log of the probability of that concept (based on the observed frequency counts).
- $IC(c) = -\log P(c)$
- Content measures of similarity can only be applied to pairs of nouns or pairs of verbs.
- Following methods are the well known Hybrid Methods;
- Resnik's Information-based Approach (1995)
- Lin's Universal Similarity Measure (1998)
- Jiang and Conrath's Combined Approach (1997)

RESNIK'S INFORMATION-BASED APPROACH (1995)

The key idea underlying Resnik's (1995) approach is the intuition that one criterion of similarity between two concepts is "the extent to which they share information in common".

- $IC(c) = -\log P(c)$
- $\text{sim}_R(c_1, c_2) = IC(lcs(c_1, c_2))$

lcs is the lowest common subsumer, it is the common concept that both c_1 and c_2 are hyponym of it.

LIN'S UNIVERSAL SIMILARITY MEASURE (1998)

The similarity between A and B is measured by the ratio between the amount of information needed to state their commonality and the information needed to fully describe what they are.


- $SIM_L(A, B) = \frac{\log P(comm(A, B))}{\log P(descr(A, B))}$
- $SIM_L(C_1, C_2) = \frac{2 \times \log P(lso(C_1, C_2))}{\log P(C_1) + \log P(C_2)}$

JIANG AND CONRATH'S COMBINED APPROACH (1997)


Jiang and Conrath's (1997) synthesize edge and node based techniques and it is formulated as;

$$\begin{aligned} \text{dist}_{JC}(c_1, c_2) &= IC(c_1) + IC(c_2) - 2 \times IC(lso(c_1, c_2)) \\ &= 2 \log p(lso(c_1, c_2)) - (\log p(c_1) + \log p(c_2)) \end{aligned}$$

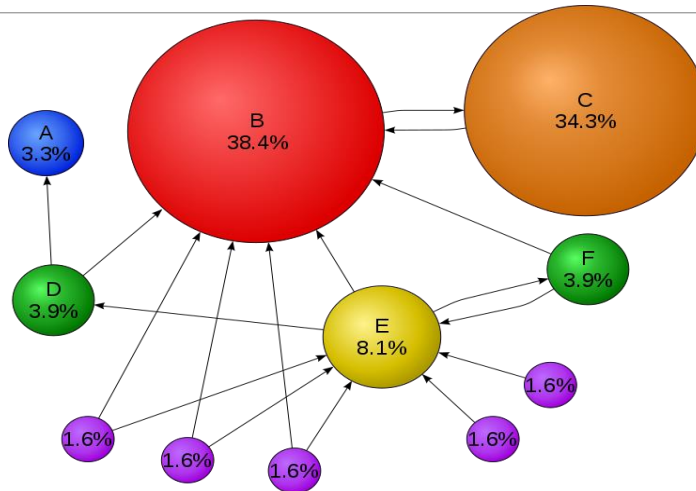
PAGERANK

- PageRank is the method applied to the graph structure to rank the each nodes in the graph.
 - It is developed by Google, it is used to rank the web url's according to their importance.
 - Importance is found using the PageRank method.
 - As a search engine Google indexes the web URLs by keywords, if there are multiple pages for that keyword, so those pages need to be ranked.
- 

PAGERANK

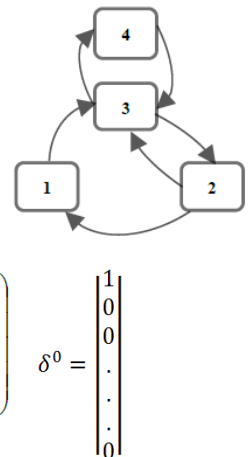
- A simple way to describe the PageRank algorithm is to consider a user who surfs the web by randomly clicking on hyperlinks.
 - The probability that the surfer will click on a specific hyperlink is given by the PageRank value of the page to which the hyperlink points.
 - PageRank also assumes that the surfer will get bored after a finite number of clicks and will jump to some other page at random.
 - The probability that our surfer will continue surfing by clicking on the hyperlinks is given by a fixed-value parameter, usually referred to as the *damping factor*..
- 

PAGERANK




PAGERANK

- PageRank generation steps;
- Consider this graph as directed graph with four nodes
- First we generate Markov Chain Matrix as below
- PageRank (δ^t in t_{th} jump) vector is generated as;
- $\delta^t = (1-\alpha) \delta^0 + \alpha M \delta^{t-1}$
- α is the dumping factor.
- Iteration is stopped after
- $|\delta^t - \delta^{t-1}| < \epsilon$ where $\epsilon < 0.0001$




$$M = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \delta^0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

PAGERANK BASED SIMILARITY

- Since the structure of the WordNet is in graph structure. PageRank can be applied to the WordNet and similarity can be calculated over PageRank vectors.
 - PageRank vectors are generated for each node in the graph, so each synset will have PageRank vectors.
 - Nodes corresponds to the web pages and relations corresponds to the links in that web page.
 - If we apply PageRank method into the WordNet, size of the PageRank vector will be 117k for each synset.
- 

PAGERANK BASED SIMILARITY

- PageRank vector for each synset is also called semantic signature of that synset.
 - We will obtain a matrix with 117K x 117K.
 - This process has been completed for WordNet 3.0 and it is available to download PageRank vectors of all of the synsets in WordNet 3.0
 - It is totally around 1.2 gb zipped data.
- 

PAGERANK BASED SIMILARITY

- Similarity computation of the PageRank vectors is the last process to get similarity score.
- There are several types of vector comparison model, but the Cosine similarity is widely used one.
- It computes the cosine value of the angle between two PageRank vectors.
- Consider A and B are the PageRank vectors of two synsets, the **Cosine** similarity is calculated as.

PAGERANK BASED SIMILARITY

- By using cosine similarity we can calculate the similarity of the given synset's by using their PageRank vectors.
- There are other comparison methods to get similarity from the PageRank vectors.
- **Jensen–Shannon Divergence Method**, This measure is based on the Kullback–Leibler divergence, which is commonly referred to as KL divergence, and is computed for a pair of semantic signatures (probability distributions in general) S1 and S2 as:

$$D_{KL}(S_1 \| S_2) = \sum_{h \in H} \log_e \left(\frac{S_1^h}{S_2^h} \right) S_1^h$$

PAGERANK BASED SIMILARITY

KL divergence is non-symmetric, so Jensen–Shannon (JS) divergence, which is a symmetrized and smoothed version of KL divergence is used:

$$D_{JS}(S_1, S_2) = \frac{1}{2} D_{KL}\left(S_1 \parallel \frac{S_1 + S_2}{2}\right) + \frac{1}{2} D_{KL}\left(S_2 \parallel \frac{S_1 + S_2}{2}\right)$$



PROJECT

Preparing a WordNet based similarity tool.

- Download the WordNet data
 - Convert it into a graph database (we suggest Neo4j)
 - Choose a similarity measure
 - Prepare a user interface
- 