NATURAL LANGUAGE PROCESSING

LESSON 9 : SEMANTIC SIMILARITY

1

OUTLINE

- Semantic Relations
- Semantic Levels
  - Sense Level
  - Word Level
  - Text Level
- Semantic Similarity Methods (Sense Level)
  - WordNet-based Similarity
  - SemSpace

2

SEMANTIC RELATIONS

- Unlike lexical similarity, semantic similarity is based on the affinity of the semantic content of the textual elements.
- There are many semantic relation types. The most important semantic relations are **synonym** and **antonym**.
- But some entities may also be semantically related by other relationships such as **meronym**, **hyponym**, **hypernym**.
  - finger is meronym of hand
  - eagle is hyponym of bird
  - bird is hypernym of eagle

3

SEMANTIC RELATIONS

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

4

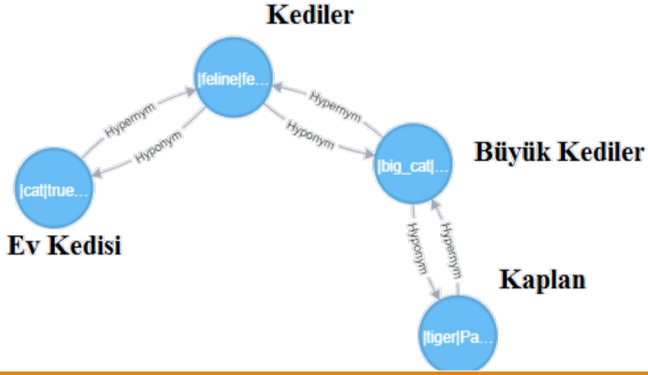
SEMANTIC RELATIONS

In fact, calculating semantic similarity and conceptual similarity between terms becomes easier when you have a dictionary.

Example:

«tiger» «house cat»

«kaplan» «ev kedisi»



5

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, the relation of related can be used (pen-paper, penguin-Antarctica, rain-flood).

6

SEMANTIC 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.

7

SENSE LEVEL

- 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.

8

SENSE LEVEL

- Sense-level semantic similarity are mostly based on dictionary or thesaurus.
- These resources are mostly used in form of semantic networks.
- In order to determine semantic similarity of two words, it is used their neighborhood.
- The most popular lexical resource is the **WordNet**.

9

SENSE LEVEL

In addition to WordNet, other resources:

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

10

WORD LEVEL

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

- Distributional approaches
- Lexical resource-based approaches

11

WORD LEVEL

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 the statistical methods such as TF–IDF.
- The dimensionality of the resulting weights matrix is often reduced, for instance using Singular Value Decomposition.
- Dictionary-based structured text content such as Wikipedia has been the source of many studies in this manner.

12

TEXT LEVEL

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.

13

TEXT LEVEL

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}

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

14

TEXT LEVEL

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

- Vector models such as TF-IDF and Document-term matrix are examples of this category.
- On the other hand, doc2vec approaches where word models such as word2vec focus on large documents have also made a significant improvement.
- In particular, transformer-based new generation contextual text vectors such as BERT and GPT achieve very successful results.

15

SEMANTIC SIMILARITY METHODS

- Sense Level
  - WordNet based methods
  - SemSpace
- Word Level
  - Word2vec ✓
- Text Level
  - TF-IDF ✓
  - Document-term matrix ✓

They can also be managed by Lexical methods

16

WORDNET BASED SIMILARITY METHODS

- WordNet is the most common structural dictionary resource and organized hierarchically in graph structure.
- It consists of nodes and edges. Nodes represent **synsets** and edges represent **relations**.
- WordNet based first methods use Hypernym, Meronomy and Antonomy relations.
- The current version of WordNet has more than 20 defined relationship types.

17

WORDNET BASED SIMILARITY METHODS

- These methods use graph structure of the WordNet, and measures similarity using several metrics such as path length, depth length, lowest common subsumer, direction of the relations.
- The following methods are the first WordNet Based similarity methods.
  - Wu & Palmer Method (1994)
  - Hirst & St-Onge Method (1998)
  - Leacock & Chodorow Method (1998)

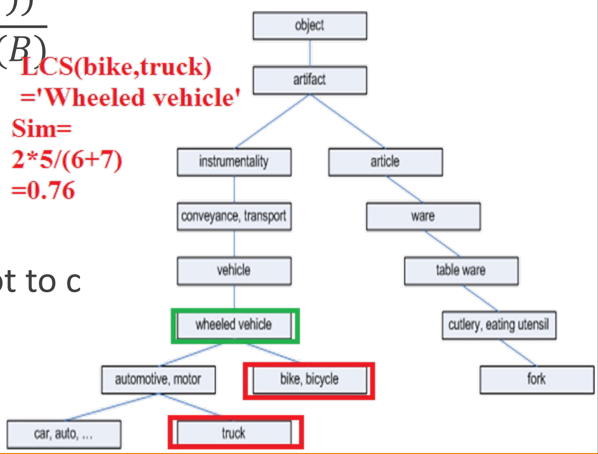
18

WU & PALMER METHOD

$$SIM_{WP}(A,B) = \frac{2 * depth(lcs(A,B))}{depth(A) + depth(B)}$$

lcs(a, b): lowest common subsumer

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



19

HIRST & ST-ONGE METHOD

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

relHS(*c*1, *c*2) = *C* – len(*c*1, *c*2) – *k* × turns(*c*1, *c*2)

where *C* and *k* are constants (in practice, they used *C* = 8 and *k* = 1), turns(*c*1, *c*2) is the number of times the path between *c*1 and *c*2 changes direction.

relHS(bike, truck) = 8-len(bike,truck)-change_of_direction

relHS(bike, truck) = 8 - 3 - 1 = 4

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

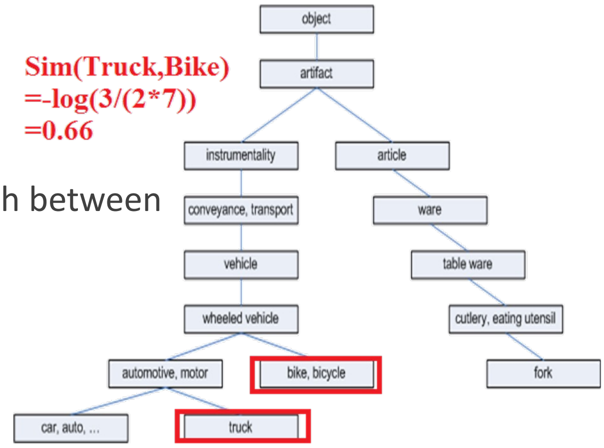
20

LEACOCK & CHODOROW METHOD

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

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

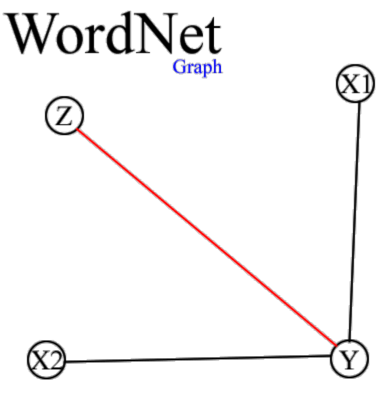
D_{max} : max depth of the taxonomy



21

SemSpace

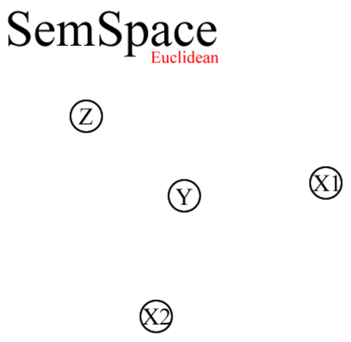
The SemSpace method aims to represent the relationships between concepts in Euclidean space by using WordNet data, which has a strong semantic graph network. With this manner, it has an approach that converts each semantic relation into distance with a special weighting method.



22

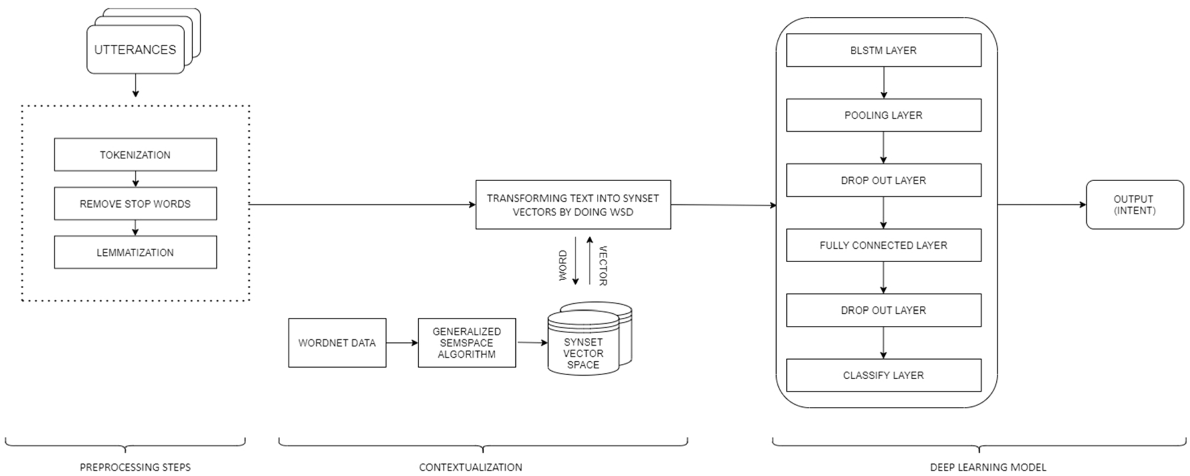
SemSpace

Thus, each concept is transformed into a physical vector representation, and used with deep learning methods to solve different NLP issues. Such vectors are particularly used in architectures known as transfer learning. By fine-tuning the pre-trained SemSpace vector model with deep learning, it can achieve higher success.



23

SemSpace



24

SUMMARY

That's all.

Please write your summary about the lesson.

25