UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Lecture

Knowledge-based Systems

Part 6 – Similarity

Dr. Mohsen Mesgar

Language Technology Lab
Universität Duisburg-Essen

Recall …

Levels of knowledge representations

Symbolic

Embeddings

Any other open questions?

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

In this lecture, you learn about …

- **Similarity**
  - in representations
  - by asking humans
- **Similarity as an evaluation methodology**
  - Correlation
  - Inter-rater agreement

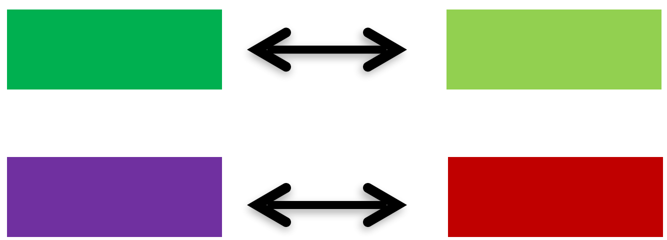
UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Similarity

Similarity between concepts

Choice of representation has large influence on similarity decisions



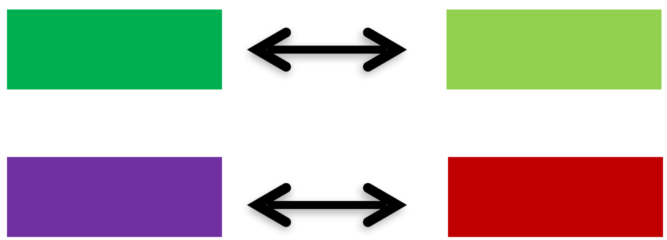
Similarity between symbols cannot be automatically determined:

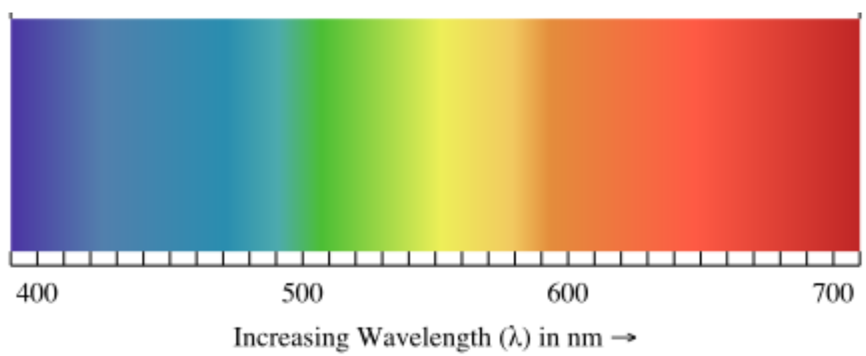
- ◊ Red, pink, rose, ruby, burgundy, magenta

- ◊ It needs to be explicitly modeled.

Similarity between concepts

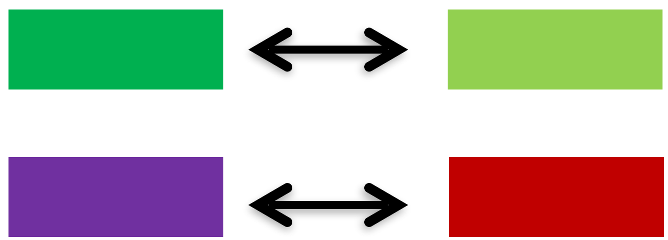
Choice of representation has large influence on similarity decisions

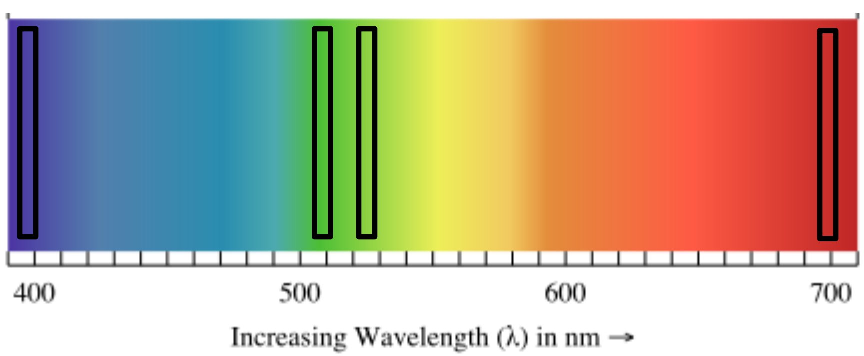




Similarity between concepts

Choice of representation has large influence on similarity decisions



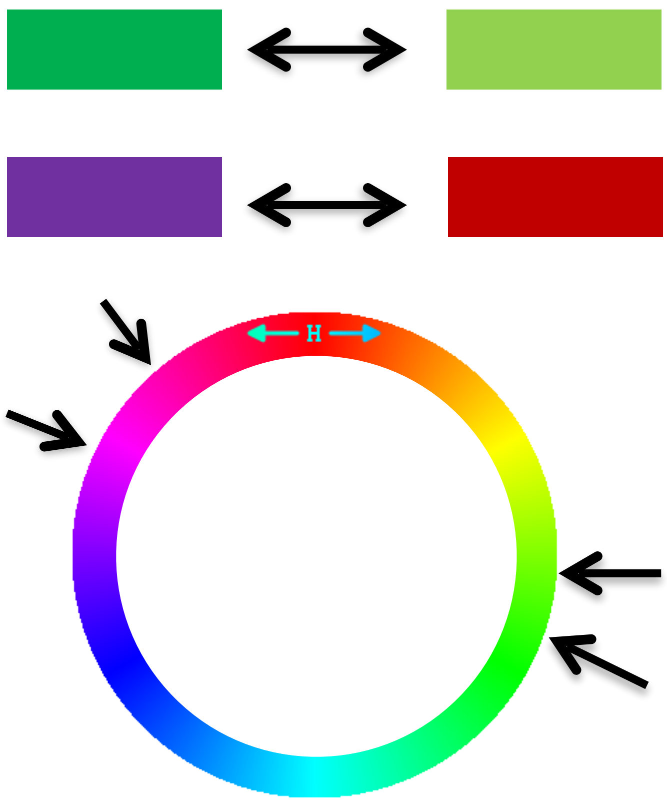


Similarity between concepts

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Choice of representation has large influence on similarity decisions



Similarity as quality

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

- The quality of knowledge representations is often evaluated by how well they model human-perceived similarity between concepts.

- How do humans perceive similarity?

Similarity Resources

- **Similarity Datasets:** humans rate pairs of words on a similarity scale
  - WordSim353 (Finkelstein et al., 2002)
  - SimLex-999 (Hill et al., 2015)

- Similarity vs Relatedness

Pair	Simlex-999 rating	WordSim-353 rating
coast - shore	9.00	9.10
clothes - closet	1.96	8.00

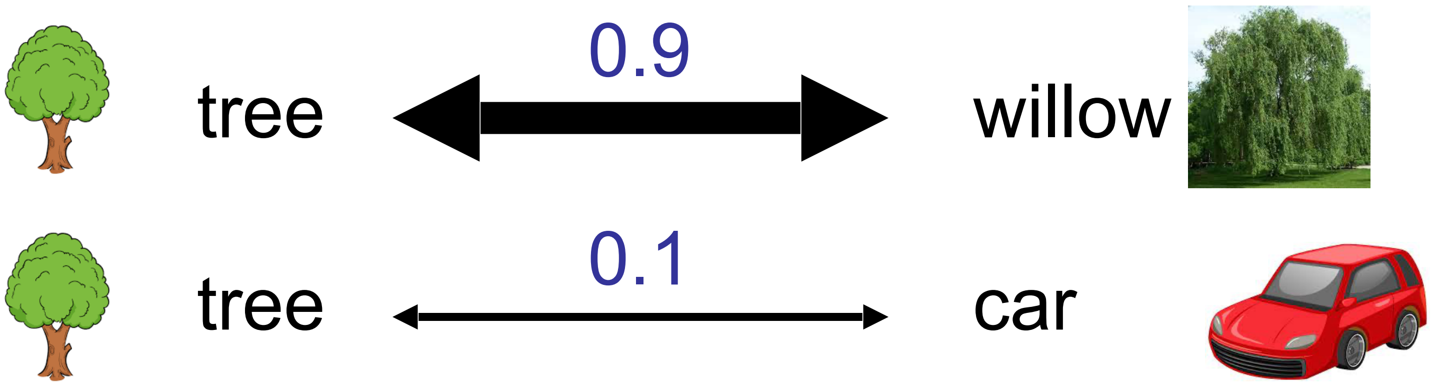
Semantic Relatedness Measures

 tree willow 

 tree car 

- Quantify the strength of semantic relatedness [0,1]

Semantic Relatedness Measures



- Quantify the strength of semantic relatedness [0,1]

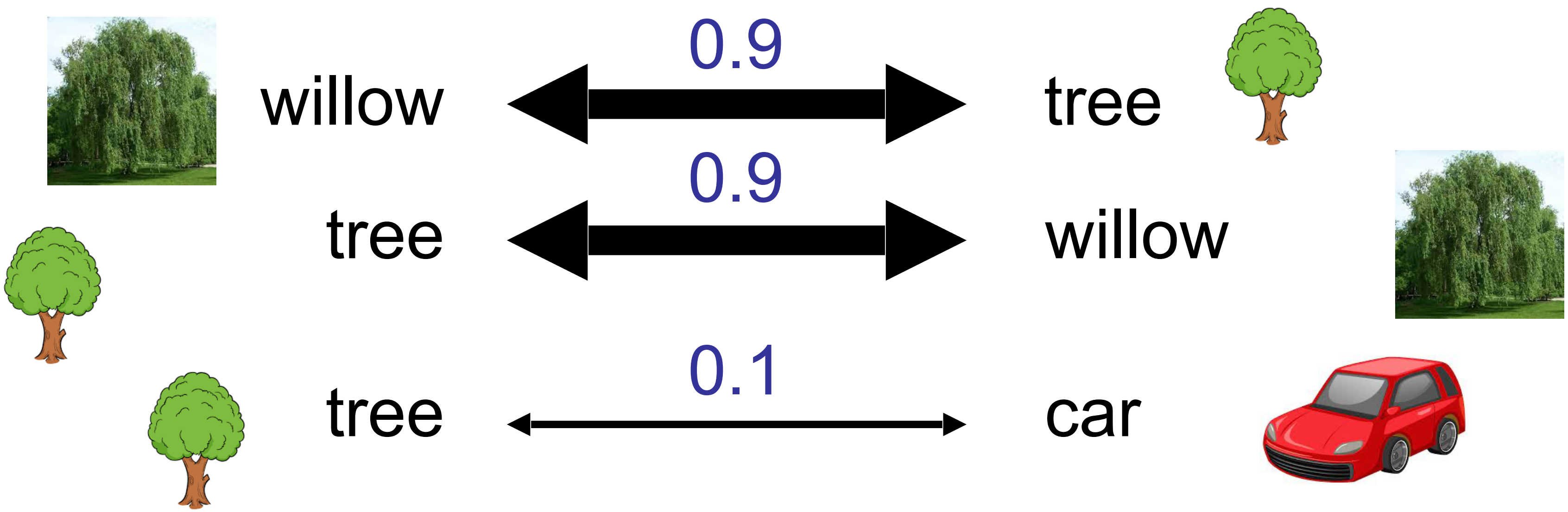
Semantic Relatedness Measures

 tree ⟷ 0.9 ⟷ willow 

 tree ⟷ 0.1 ⟷ car 

- Quantify the strength of semantic relatedness [0,1]
- Symmetric

Semantic Relatedness Measures

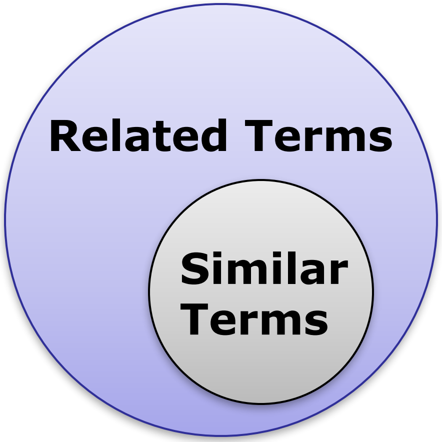


- Quantify the strength of semantic relatedness [0,1]
- Symmetric

Semantic Similarity vs. Semantic Relatedness

Semantic Relatedness is broader than similarity

- Concepts can be related without being similar
  - car / automobile vs. car / street
- Defined across different parts of speech
  - car / drive
  - night / dark
- Defined for antonyms
  - night / day



Semantic Relatedness & Senses

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

How related are `bat` and `ball`?

Semantic Relatedness & Senses

How related are `bat` and `ball`?

- Highly related

Semantic Relatedness & Senses

How related are `bat` and `ball`?

- Highly related

How related are `wing` and `bat`?

Semantic Relatedness & Senses

How related are `bat` and `ball`?

- Highly related

How related are `wing` and `bat`?

- Highly related

Semantic Relatedness & Senses

How related are **`bat`** and **`ball`**?

- Highly related

How related are **`wing`** and **`bat`**?

- Highly related

Humans implicitly assume the sense that maximizes relatedness

Semantic Relatedness & Senses

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

bat ⟷ ball

bat ⟷ wing

From Senses to Words

Semantic relatedness of words is the maximum semantic relatedness value between all concept pairs

bat

ball

bat

$$rel = \begin{cases} \min\limits_{c_1 \in C(w_1), c_2 \in C(w_2)} & dist(c_1, c_2) \\ \max\limits_{c_1 \in C(w_1), c_2 \in C(w_2)} & rel(c_1, c_2) \end{cases}$$

Similarity datasets to evaluate knowledge representations

- Similarity datasets are often used to evaluate knowledge representations
- For example:
  - correlation between similarity values and graph-based distance in resource
  - correlation between similarity values and gloss overlap in dictionary

Semantic Relatedness Measures

We can categorize these measures by knowledge source:

Symbolic Knowledge bases

- Using a knowledge source with explicit sense distinctions

Distributional

- Using the distributional context of words in a corpus

Semantic Relatedness Measures

We can categorize them by method:

Path based

- Relying on paths in a structured knowledge source

Definition based

- Comparing glosses

Vector based

- Relying on vector representations of words

Outline

Measures of Semantic Relatedness

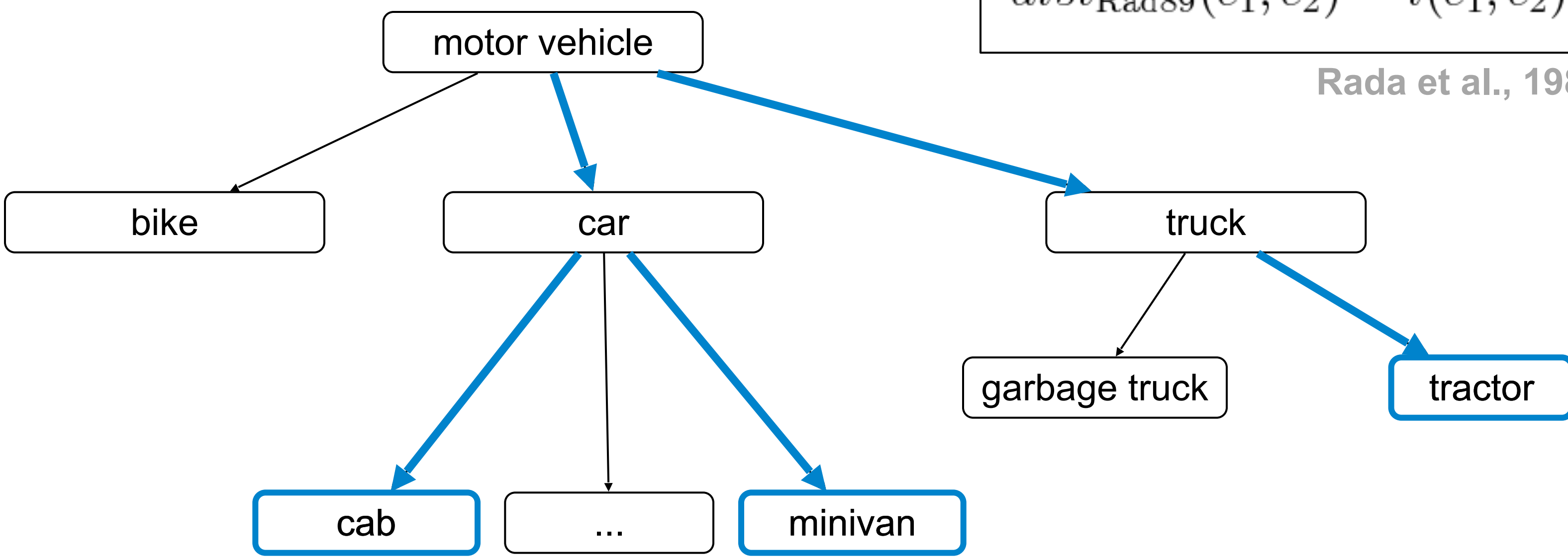
- Path based
- Definition based
- Vector based

Path Based Measures

Semantic relatedness corresponds e.g. to number of edges of the shortest path between two nodes in graphical knowledge bases (articles, categories)

$$dist_{\mathrm{Rad89}}(c_1, c_2) = l(c_1, c_2)$$

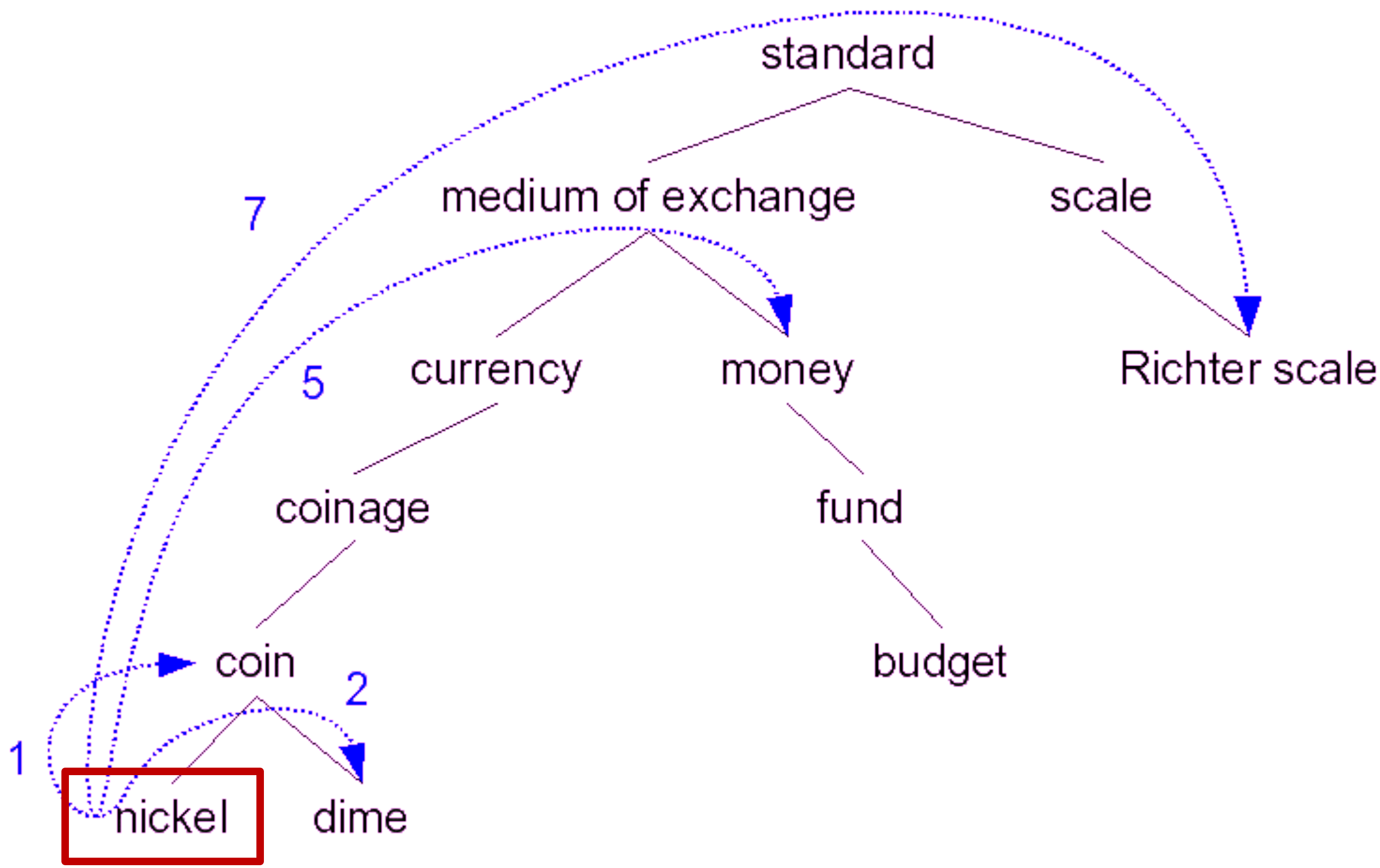
Rada et al., 1989



cab – minivan: 2

cab – tractor: 4

Path Based Measures

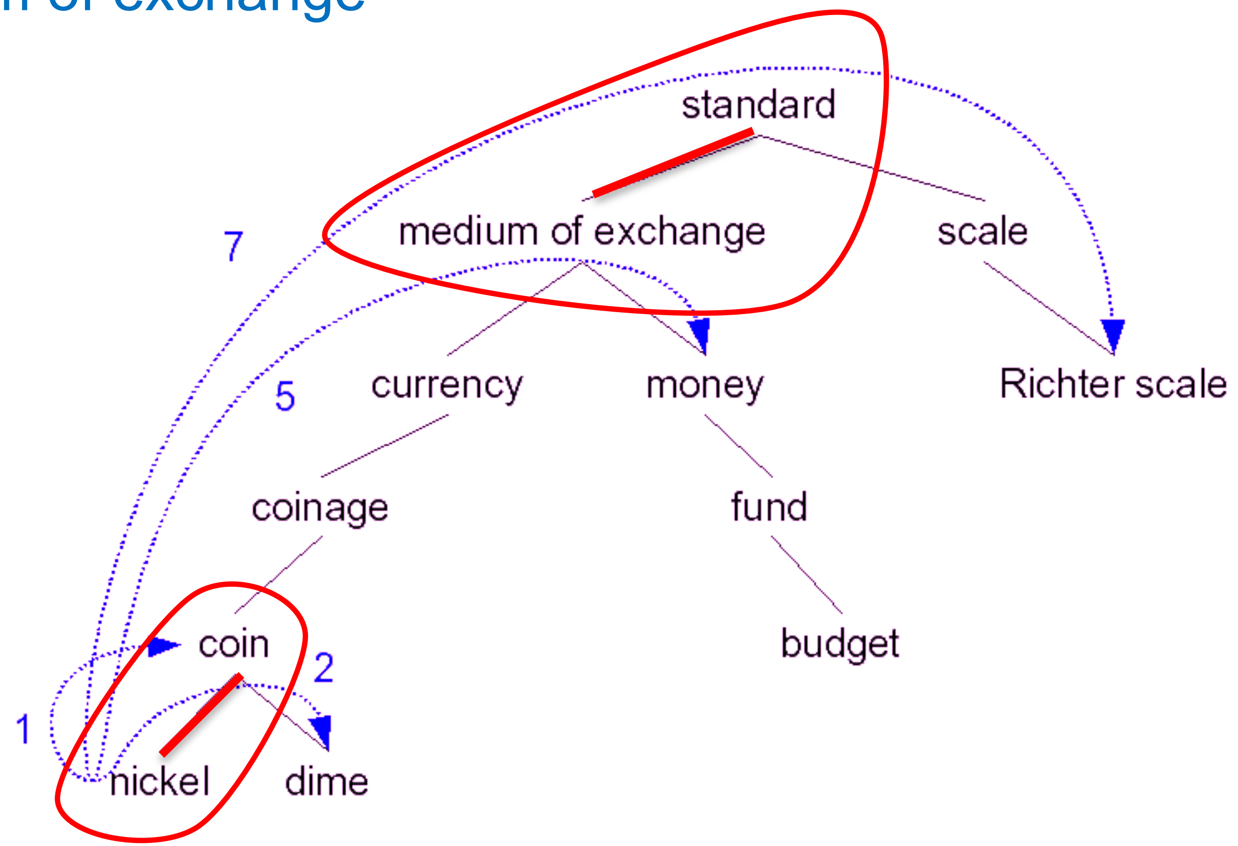


Jurafsky & Martin

Issues with Path Based Relatedness

Links do not express uniform distance

- standard / medium of exchange
- nickel / coin



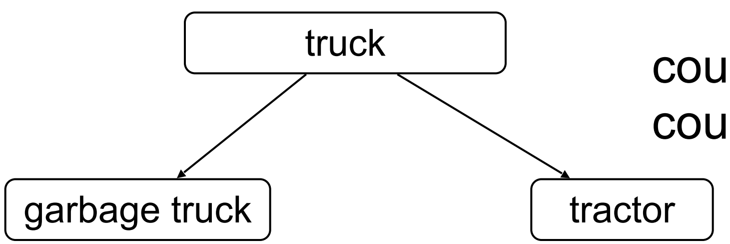
Information Content

The probability that a word in a corpus is an instance of concept *c*

- IC = - log(p(c))
- p(root) = 1 → IC(root) = 0

Estimated as

$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$$



count(truck) →
count(truck) + count(garbage truck) + count(tractor)

The lower a node in the hierarchy is, the higher is its information content

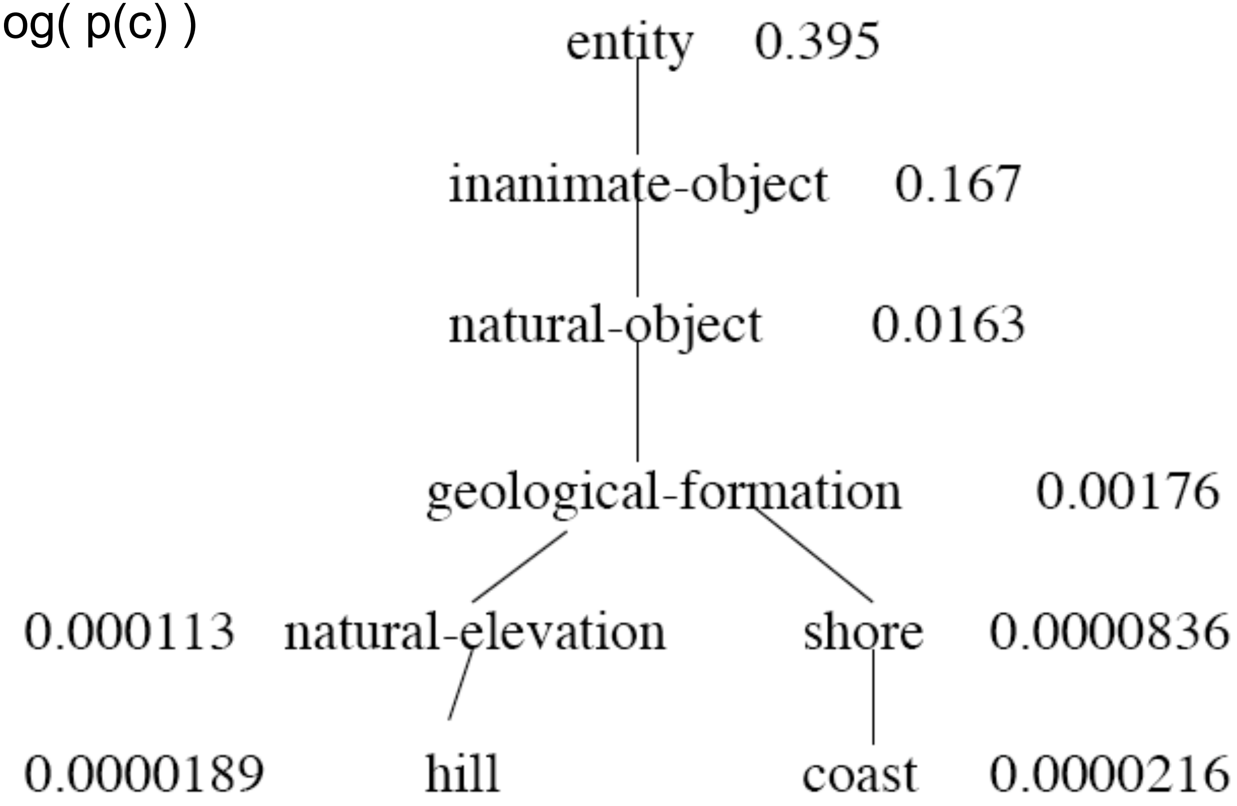
Information Content Example

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

WordNet hierarchy augmented with probabilities p(C)

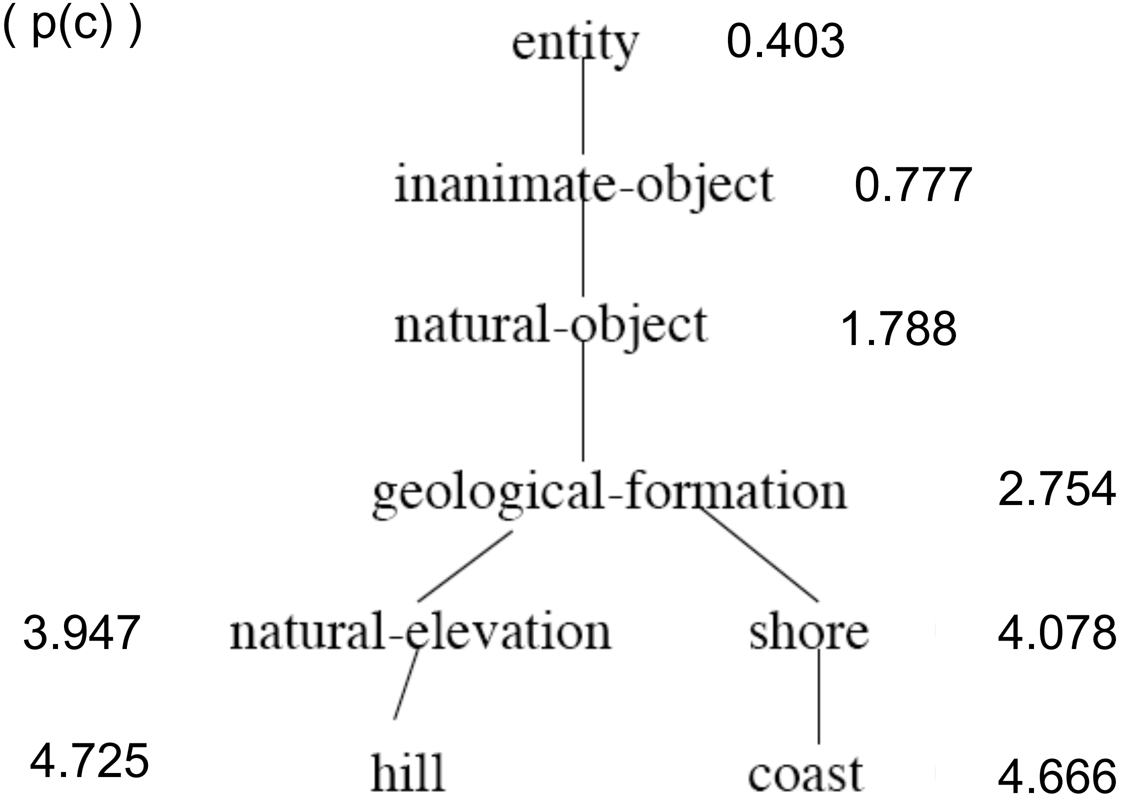
IC = - log(p(c))



Information Content Example

WordNet hierarchy augmented with Information Content

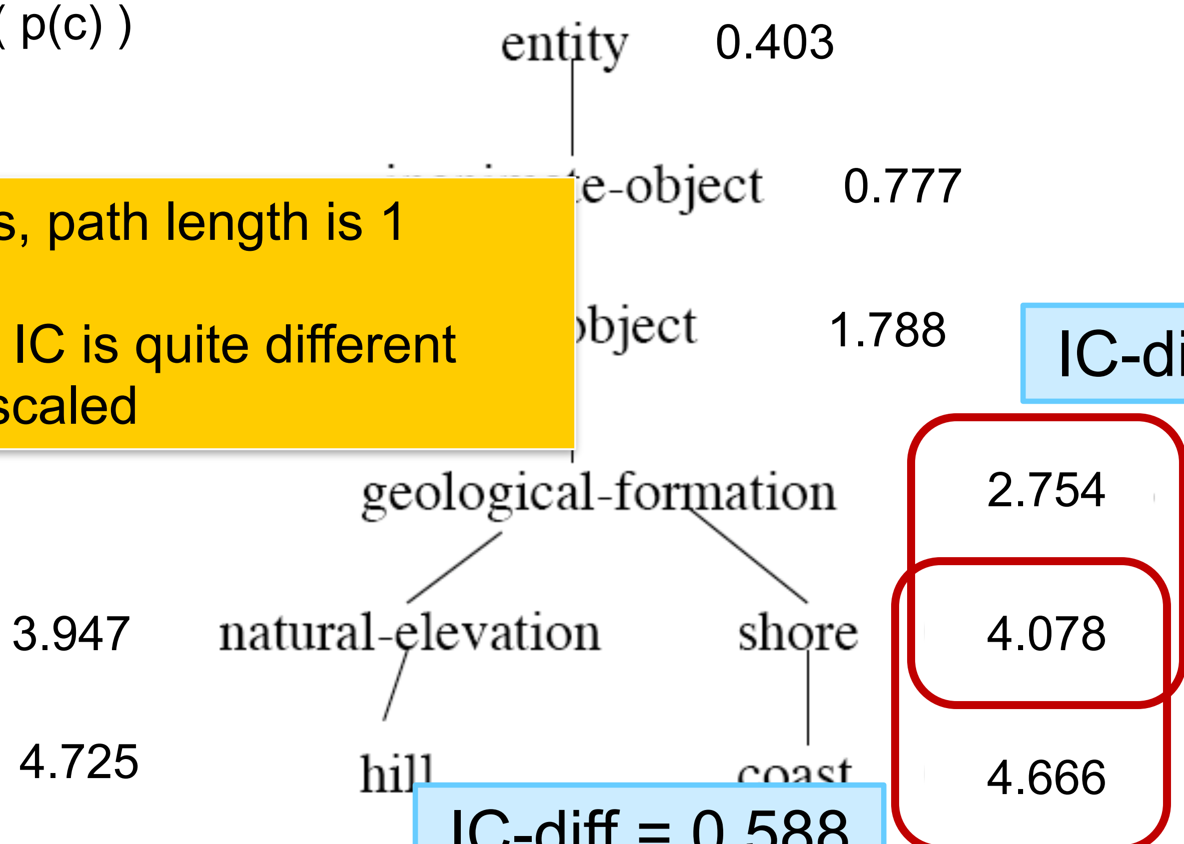
IC = - log(p(c))



Information Content Example

WordNet hierarchy augmented with Information Content

IC = - log(p(c))



In both cases, path length is 1

Difference in IC is quite different

- Paths are scaled

IC-diff = 1.324

IC-diff = 0.588

Practice

- Make a group of 2 or 3
- Load the following notebook into Google Colab
  - https://github.com/MMesgar/Knowledge_Based_Systems/blob/main/lecture06/semantic_relatedness.ipynb
- Complete the first section

Outline

Introduction to Semantic Relatedness

Measures of Semantic Relatedness

- Path based
- **Definition based**
- Vector based

Evaluation

Gloss Based Measures

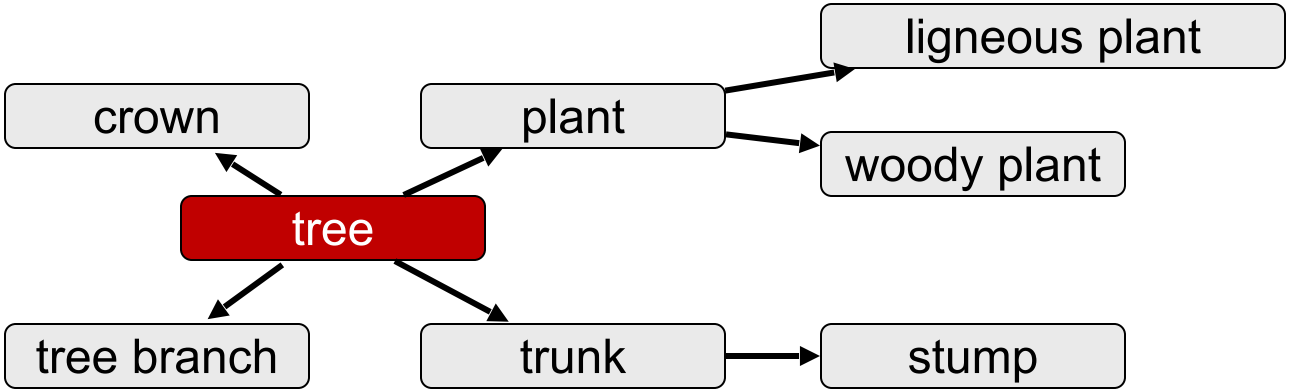
UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

WordNet glosses

- tree (plant) "a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown"
- trunk (tree) "the main stem of a tree; usually covered with bark; the bole is usually the part that is commercially useful for lumber"

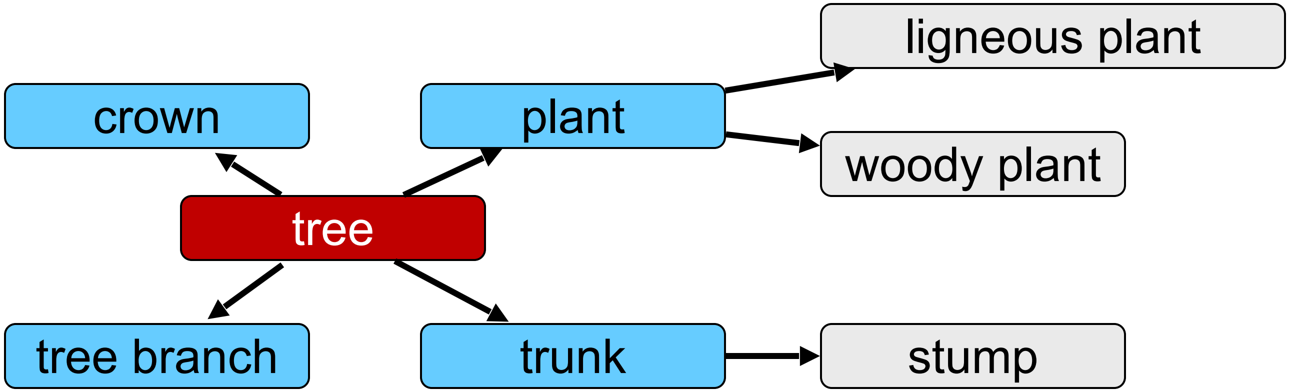
Substituting Glosses with Pseudo Glosses



- Pseudo gloss
  - tree

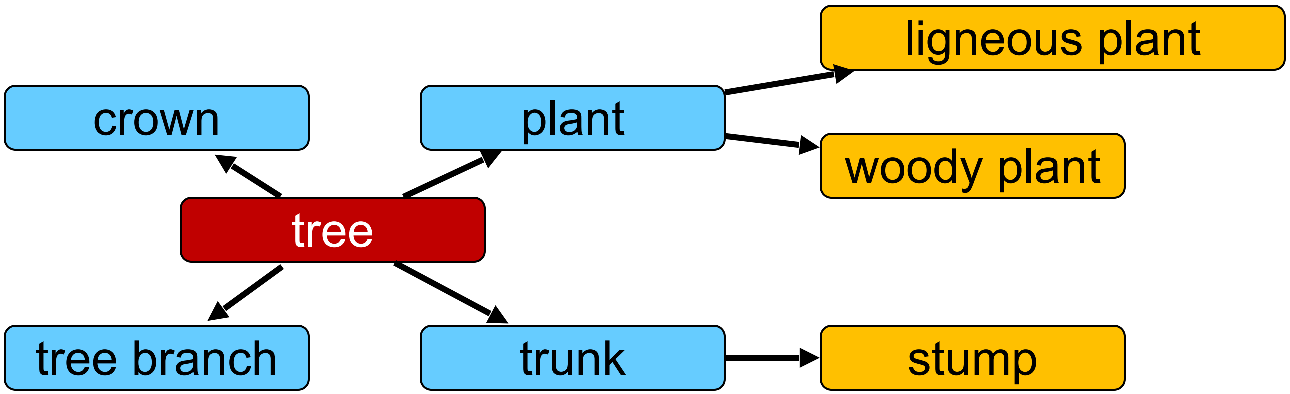
- WordNet gloss
  - tree (plant) “a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown”

Substituting Glosses with Pseudo Glosses



- Pseudo gloss
  - tree, plant, crown, branch, trunk
- WordNet gloss
  - tree (plant) “a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown”

Substituting Glosses with Pseudo Glosses



- Pseudo gloss
  - tree, plant, crown, branch, trunk, ligneous, woody, stump

- WordNet gloss
  - tree (plant) "a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown"

Term – Document Matrix

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Terms

	t_1	t_2	t_3	…	t_{m-1}	t_m
d_1	3	1	0	…	0	0
d_2	0	5	0	…	1	0
d_3	1	0	2	…	3	3
…	…	…	…	…	…	…
d_{n-1}	0	2	3	…	2	1
d_n	2	3	0	…	5	0

Documents

Term – Document Matrix

	t_1	t_2	t_3	...	t_{m-1}	t_m
d_1	3	1	0	…	0	0
d_2	0	5	0	…	1	0
d_3	1	0	2	…	3	3
…	…	…	…	…	…	…
d_{n-1}	0	2	3	…	2	1
d_n	2	3	0	…	5	0

Definition Based Measures

		t_1	t_2	t_3	...	t_{m-1}	t_m
c_1 →	d_1	3	1	0	...	0	0
c_2 →	d_2	0	5	0	...	1	0
c_3 →	d_3	1	0	2	...		
	...	...	...	...	...	...	...
c_{n-1} →	d_{n-1}	0	2	3	...	2	1
c_n →	d_n	2	3	0	...	5	0

Inner Product
(usually Lesk)

Concepts

Glosses

$$rel_{\text{Les86}}(c_1, c_2) = |gloss(c_1) \cap gloss(c_2)|$$

[Lesk, 1986]

Practice

- Make a group of 2 or 3
- Load the following notebook into Google Colab
  - https://github.com/MMesgar/Knowledge_Based_Systems/blob/main/lecture06/semantic_relatedness.ipynb
- Complete the second section

Outline

Introduction to Semantic Relatedness

Measures of Semantic Relatedness

- Path based
- Definition based
- Vector based

Evaluation

Wikipedia Link Vector Measure

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Links

		d1	d2	d3	…	d4	d5
c_1	d1	-	1	0	…	0	0
c_2	d2	0	-	1	…	1	0
c_3	d3	1	0	-	…		
	…	…	…	…	…	…	…
c_{n-1}	d_{n-1}	0	2	3	…	-	1
c_n	d_n	2	3	0	…	5	-

Inner Product (usually Cosine)

Article Titles

Articles

$$\omega(s \rightarrow t) = \begin{cases} \log\left(\frac{N}{|T|}\right), s \in T \\ 0,\ otherwise \end{cases}$$

https://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-005.pdf

Drawbacks of symbolic knowledge based similarity measures

Not every language has large knowledge bases

They are often quite incomplete

Alternative:

- Distributional approach

Distributional Approach

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

[Firth, 1957]

- “You shall know a word by the company it keeps.”

[Lin, 1998] after [Nida, 1975]

- A bottle of ***tezgüino*** is on the table
- Everybody likes ***tezgüino***
- ***Tezgüino*** makes you drunk
- We make ***tezgüino*** out of corn.

What is “tezgüino”?

Intuition:

- Just from these contexts, a human could guess the meaning of tezgüino.
- Thus, we should look at the surrounding contexts, see what other words have similar context.

Distributional hypothesis

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- perhaps we can infer meaning just by looking at the contexts a word occurs in
- perhaps meaning IS the contexts a word occurs in (Wittgenstein!)
- either way, similar contexts imply similar meanings:
  - this idea is known as the distributional hypothesis

Term Document Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Term Document Matrix
- Context = Document in which a term occurs

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

TF.IDF Term Weighting

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- common words influence the similarity too much
  - "The king is here." vs. "The salad is cold."
  - solution: weigh words for "importance"

TF(w,d) – Term Frequency

- the number of times word *w* occurs in document *d.*

DF(w) – Document Frequency

- the number of documents in which the word *w* occurs

IDF(w) – Inverse Document Frequency

$$IDF(w) = \log\left(\frac{|D|}{DF(w)}\right)$$

- Largest value for terms that occur often in a certain document, but cannot be found in other documents

Sparse Vectors

	[...] aardvark	computer	data	pinch	result	sugar [...]
apricot	0	0	0	1	0	1
pineapple	0	0	0	1	0	1
digital	0	2	1	0	1	0
information	0	1	6	0	4	0

- What about the word lemon?

- shown matrix is 4x6
- the real matrix is much larger (e.g. 50,000 x 50,000)
- sparse matrix (most values are 0)

Term Document Matrix

- is a document expressive context?
- instead of entire documents, use smaller contexts
- variants:
  - paragraph
  - window of ± N words
  - syntactic context
  - …

Word-context Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- we count how often a word has occurred with a context in texts from a large corpus
- context is defined by a window over words
- let V be the set of words in vocabulary and C be the set of possible contexts
- word-context matrix M is a two dimensional matrix whose rows are associated with V and columns with C
- each entry of the matrix indicates how often a word co-occurs with a context in the given corpus
- this matrix is also known as co-occurrence matrix

Word-context Matrix

- corpus:
  - I like DL
  - I like NLP
  - I love ML
  - I love NLP
- window size = 1

	I	like	love	DL	NLP	ML
I	0	2	2	0	0	0
like	2	0	0	1	1	0
love	2	0	0	0	1	1
DL	0	1	0	0	0	0
NLP	0	1	1	0	0	0
ML	0	0	1	0	0	0

taken from my course: DL4NLP

Word-context Matrix

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- corpus:
  - I like DL
  - I like NLP
  - I love ML
  - I love NLP
- window size = 1

	I	like	love	DL	NLP	ML
I	0	2	2	0	0	0
like	2	0	0	1	1	0
love	2	0	0	0	1	1
DL	0	1	0	0	0	0
NLP	0	1	1	0	0	0
ML	0	0	1	0	0	0

taken from my course: DL4NLP

Practice

- Make a group of 2 or 3
- Load the following notebook into Google Colab
  - https://github.com/MMesgar/Knowledge_Based_Systems/blob/main/lecture06/semantic_relatedness.ipynb
- Complete the third section, Word-Context Matrix

Sparse vs. Dense Vectors

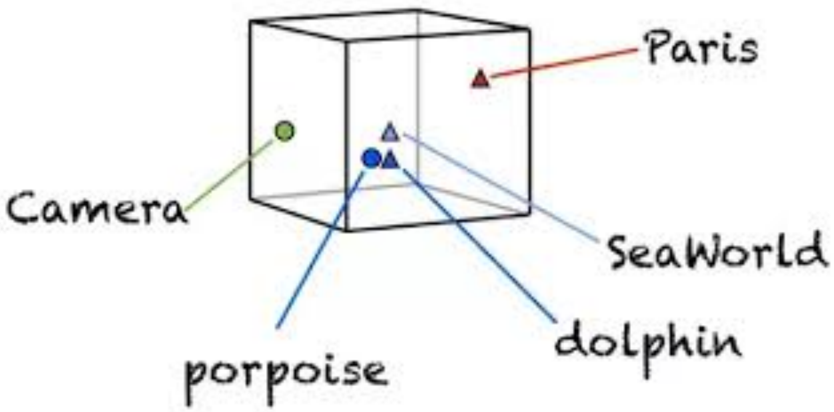
- sparse vectors are
  - long (length |V|= 20,000 to 50,000)
  - redundant (most elements are zero)

- advanced methods rely on <u>dense</u> vectors
  - short (length 25-1000)
  - dense (most elements are non-zero)

Why Dense Vectors?

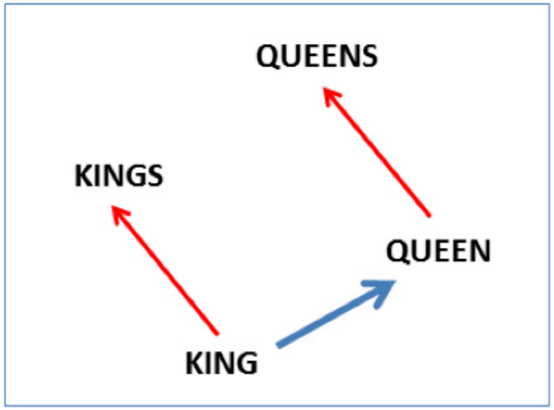
- They are not sparse
- key component of many breakthroughs in NLP

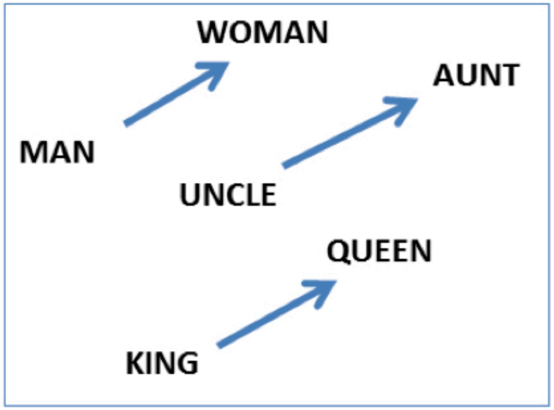
- more <u>presumed</u> advantages:
  - Denoising: low-order dimensions may represent unimportant information
  - Truncation: may help the models generalize better to unseen data
  - More Semantic: better at capturing higher order co-occurrence

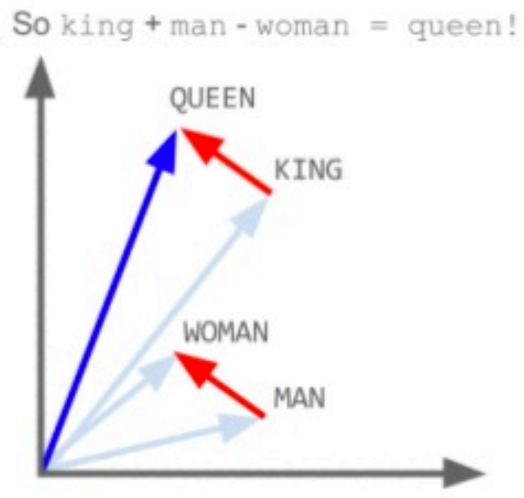


Relational meaning with dense word vectors

- dense word vectors encode rich semantics
- **but**: highly dependent on characteristics of train corpus







Practice

- Make a group of 2 or 3
- Load the following notebook into Google Colab
  - ss
- Complete the third section, Dense vectors

How to create dense word vectors

- several methods
  - decomposition of a term-context matrix
  - compression of a term-context matrix
  - clustering words based on similar contexts

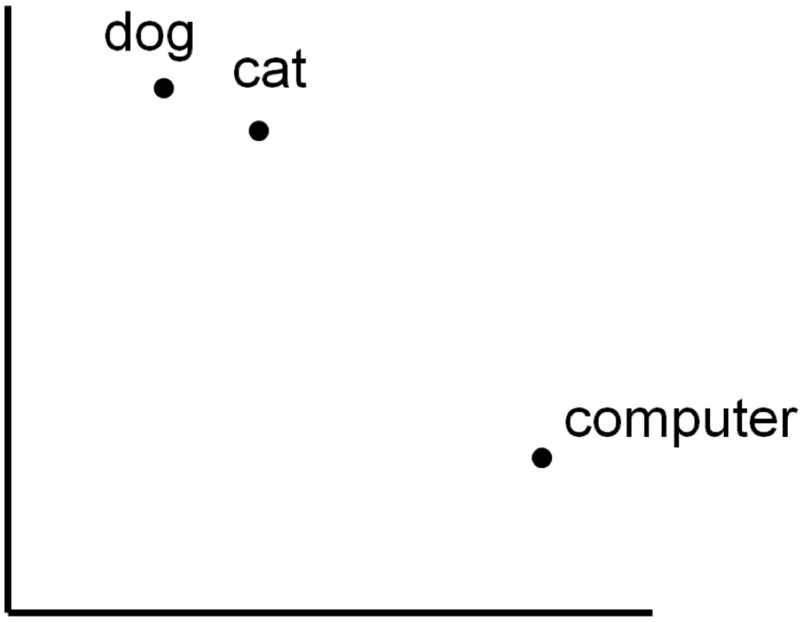
- active research field / constant new developments

How to measure similarity?

- let's assume we have context vectors for two words $\vec{v}$ and $\vec{w}$
- one way to think of these vectors: as points in high-dimensional space

How to measure similarity?

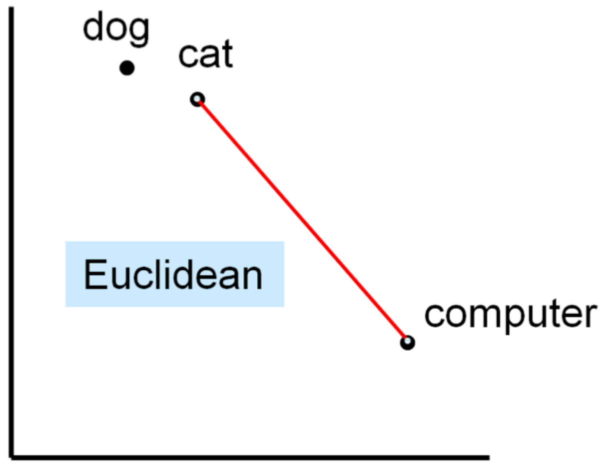
- let's assume we have context vectors for two words $\vec{v}$ and $\vec{w}$
- one way to think of these vectors: as points in high-dimensional space



How to measure similarity?

- Euclidian distance

  - We could measure (dis)similarity using Euclidean distance: $\left(\sum_i (v_i - w_i)^2\right)^{1/2}$



  - But doesn't work well if even one dimension has an extreme value

https://www.inf.ed.ac.uk/teaching/courses/fnlp/lectures/15_slides.pdf

How to measure similarity?

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

- Dot product

  - Another possibility: take the dot product of $\vec{v}$ and $\vec{w}$:

$$\begin{aligned} \text{sim}_{\text{DP}}(\vec{v}, \vec{w}) &= \vec{v} \cdot \vec{w} \\ &= \sum_i v_i w_i \end{aligned}$$

    - Vectors are longer if they have higher values in each dimension.
    - So more frequent words have higher dot products.

https://www.inf.ed.ac.uk/teaching/courses/fnlp/lectures/15_slides.pdf

How to measure similarity?

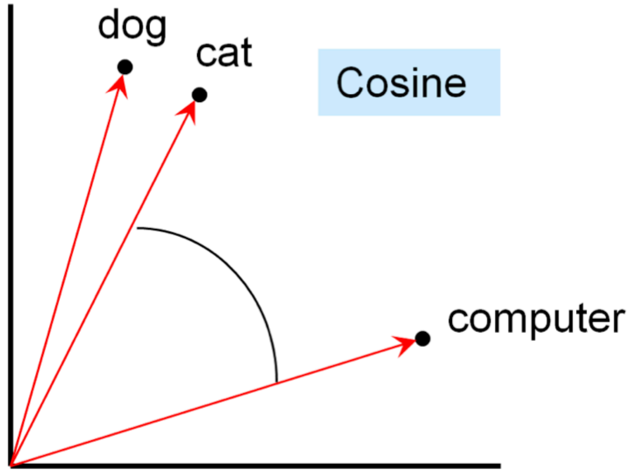
- Normalized dot product
  - divide each vector by the length of the vector
  -

$$\text{sim}_{\text{NDP}}(\vec{v}, \vec{w}) = (\vec{v} \cdot \vec{w}) / (|\vec{v}||\vec{w}|)$$

https://www.inf.ed.ac.uk/teaching/courses/fnlp/lectures/15_slides.pdf

How to measure similarity?

- The normalized dot product is just the cosine of the angle between vectors.



- Ranges from -1 (vectors pointing opposite directions) to 1 (same direction

https://www.inf.ed.ac.uk/teaching/courses/fnlp/lectures/15_slides.pdf

Practice

- Make a group of 2 or 3
- Load the following notebook into Google Colab
  - https://github.com/MMesgar/Knowledge_Based_Systems/blob/main/lecture06/semantic_relatedness.ipynb
- Complete the third section

Evaluation

What is evaluation?

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Evaluation is systematic determination of merit, worth, and significance of something or someone using criteria against a set of standards.

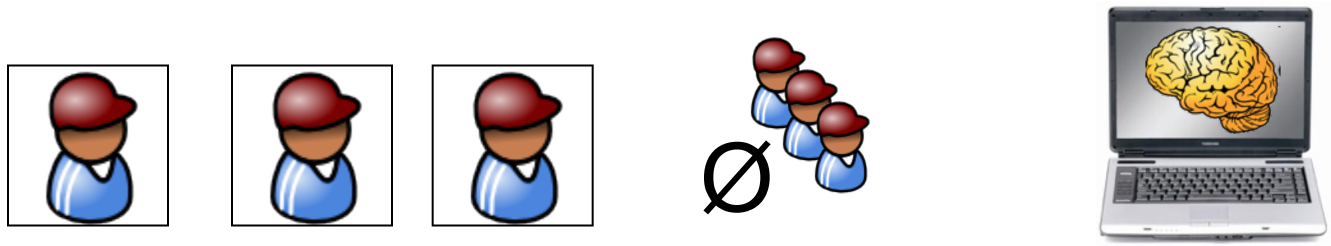
- Repeatable
- Documented

- Precision
- Recall
- Correlation
- Turing-Test
- …

- Human performance
- Baselines
- …

Definition from Wikipedia

Evaluation process



tree – lake	0.5	0.5	0.75	0.58	0.7
tree – willow	0.75	0.75	1.0	0.83	0.9
tree – car	0.25	0.0	0.0	0.08	0.0

Correlation coefficient (Pearson / Spearman)

Evaluation process

				Gold Standard	
tree – lake	0.5	0.5	0.75	0.58	0.7
tree – willow	0.75	0.75	1.0	0.83	0.9
tree – car	0.25	0.0	0.0	0.08	0.0

Correlation coefficient (Pearson / Spearman)

Criteria

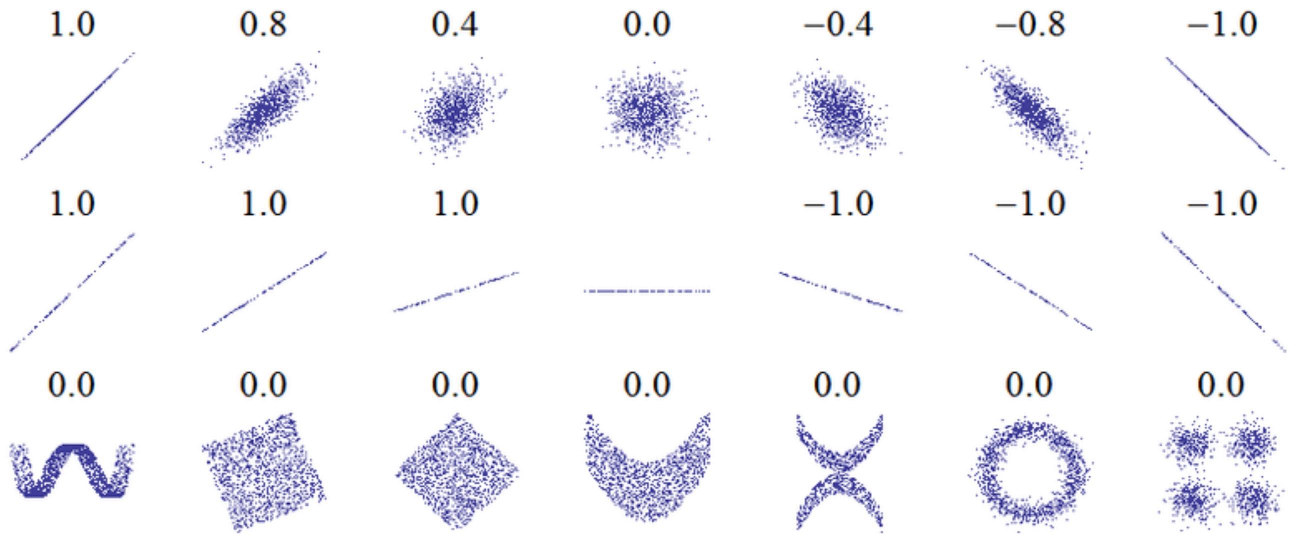
Pearson correlation r

Measure of linear statistical dependence between two variables X and Y

- Named after Karl Pearson
- Often denoted as Pearson's r

r =

$$\frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$



Pearson's r: Example

X	Y
1	2
2	5
3	6

r =

$$\frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

Practice

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Make a group of 2 or 3
- Load the following notebook into Google Colab
  - https://github.com/MMesgar/Knowledge_Based_Systems/blob/main/lecture06/semantic_relatedness.ipynb
- Complete the fourth section

Pearson Correlation r – Example

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

X	Y
1	2
2	5
3	6

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

$\Sigma XY = (1)(2) + (2)(5) + (3)(6) = 30$

$\Sigma X = 1 + 2 + 3 = 6$

$\Sigma X^2 = 1^2 + 2^2 + 3^2 = 14$

$\Sigma Y = 2 + 5 + 6 = 13$

$\Sigma Y^2 = 2^2 + 5^2 + 6^2 = 65$

$N=3$

$\Sigma XY - \Sigma X \Sigma Y/N = 30 - (6)(13)/3 = 4$

$\Sigma X^2 - (\Sigma X)^2/N = 14 - 6^2/3 = 2$

$\Sigma Y^2 - (\Sigma Y)^2/N = 65 - 13^2/3 = 8.667$

$r = 4 / \sqrt{(2)(8.6667)} = 4/4.16333$

$= .9608$

Criticism

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Look at the four sets of data

- Same mean (7.5), standard deviation (4.1), regression line (y=3+0.5x)
- Same Pearson correlation
  - 0.816
- Very different distributions

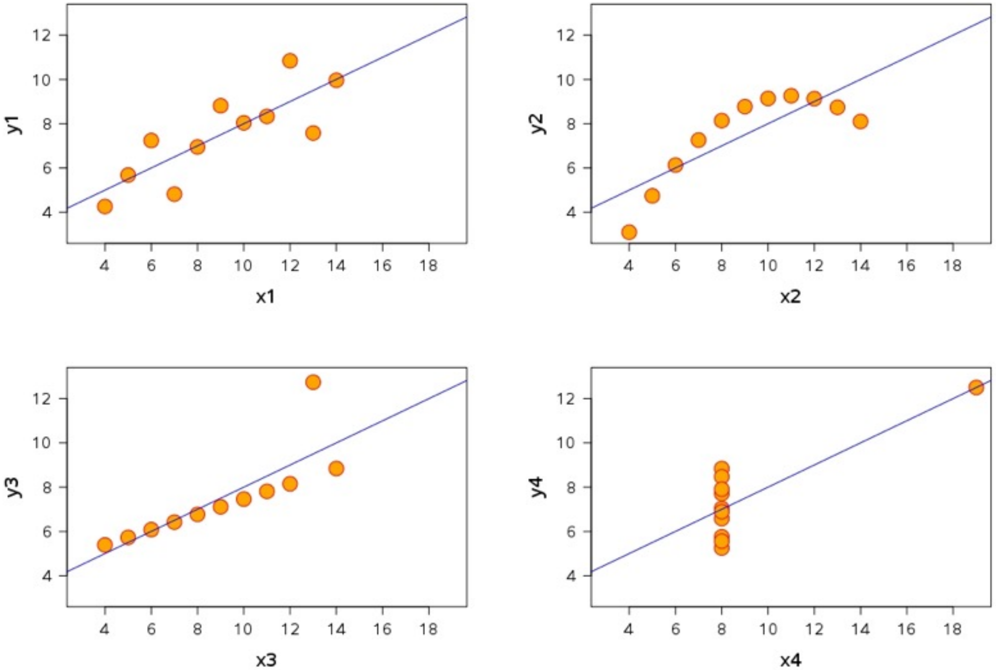


Image from Wikipedia

Spearman correlation ρ

Non-parametric measure of statistical dependence

- Named after Charles Spearman
- Often denoted by the Greek letter ρ (rho) or as r_s

Describes how well the relationship between two variables can be described using a monotonic function

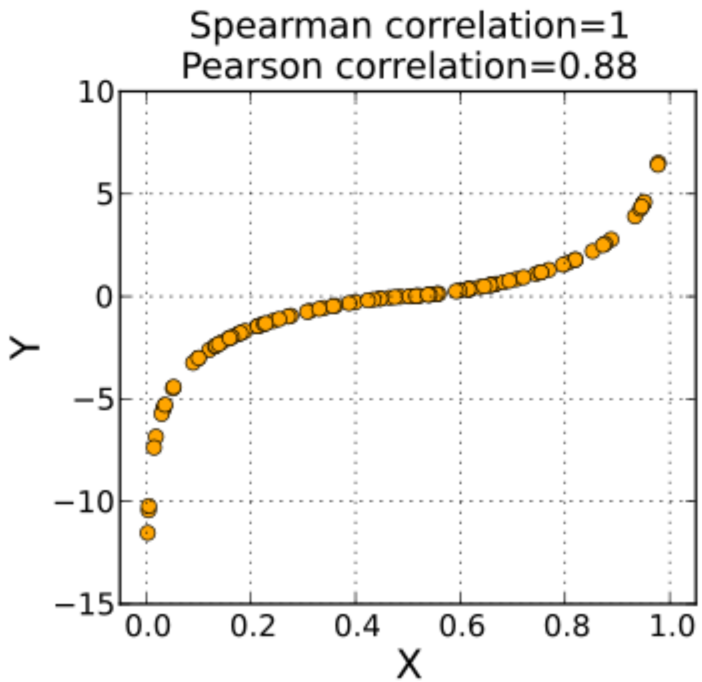
Spearman correlation is equivalent to Pearson correlation on the ranks instead of the actual values.

Spearman vs. Pearson

A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. In contrast, this does not give a perfect Pearson correlation.

Spearman: 1.00

Pearson: 0.88



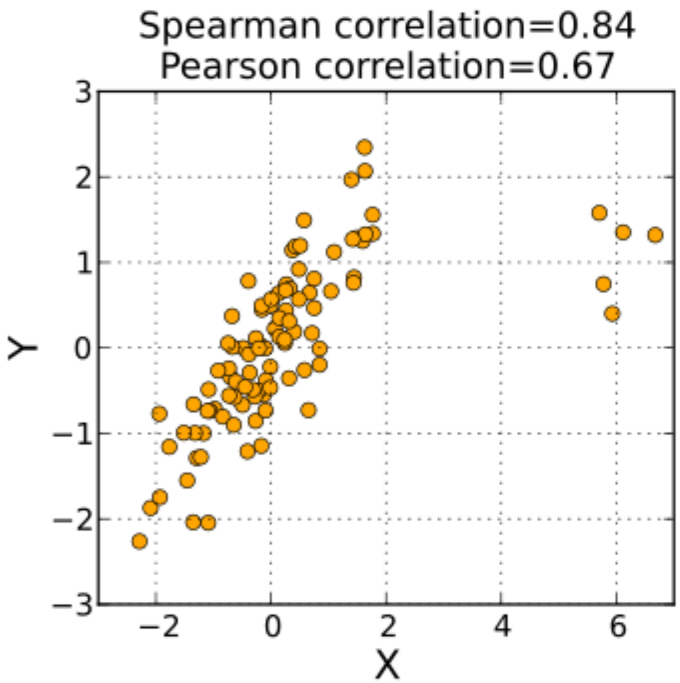
Example from Wikipedia

Spearman vs. Pearson

The Spearman correlation is less sensitive than the Pearson correlation to strong outliers.

Spearman: 0.84

Pearson: 0.67



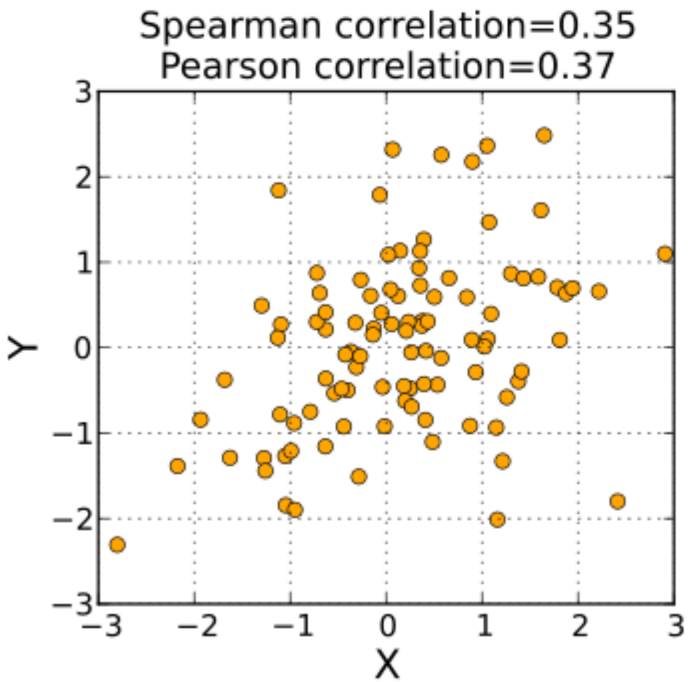
Example from Wikipedia

Spearman vs. Pearson

When there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.

Spearman: 0.35

Pearson: 0.37



Example from Wikipedia

Careful!

- Do not rely on correlation values alone!
- Always inspect the data!

- Statistics can be misleading.

Correlation vs Causality

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- The values of variable A and B are highly correlated:
- Possible, but not necessarily valid conclusion:
  → A causes B
- Other explanations:
  - B causes A:
    The faster windmills are observed to rotate, the more wind is observed.
    → ~~Windmills produce wind.~~
  - A and B are both caused by a third (unknown) variable:
    Sleeping with one's shoes on is strongly correlated with waking up with a headache. → ~~shoes cause headaches~~
  - The relationship between A and B is coincidental:
    A bald (or obviously balding) state leader of Russia has succeeded a non-bald ("hairy") one, and vice versa, for nearly 200 years.→ ??? https://en.wikipedia.org/wiki/Bald%E2%80%93hairy

Correlated features

- In knowledge-based systems, highly correlated properties are often interpreted as redundant information and the system focuses on one of them to optimize processing.

- This might lead to wrong conclusions:
  - Sleeping with shoes and being drunk are highly correlated
    → focus on sleeping with shoes as a predictor for headaches.
  - Redness and ripeness of fruits are correlated
    → focus on redness as indicator for edibility.

How golden is the gold standard?

- If we compare to human ratings, how do we now that the human ratings are of high quality?

- Humans might be:
  - tired, bored, drunk, insecure, mischievous, lazy, careless, distracted,...

- The task might be
  - ambiguous, unclear, too hard, too easy, subjective,...

Validity, Reliability, Agreement

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- For each (manually or automatically generated) dataset, it is crucial to consider the following questions:

Is my evaluation valid? → **Is my evaluation data reliable?** → **What is good agreement?**

Is my evaluation valid?

◊ Can we draw conclusions from the data?

◊ One prerequisite for validity is that the evaluation data is reliable.

Is my evaluation data reliable?

◊ Is the generation reproducible?

◊ Raters annotate a sample of the data.

◊ **Assumption:** The data is reliable if the raters reach good agreement.

What is good agreement?

◊ How to measure agreement?

◊ How to interpret the result?

◊ **Inter-rater agreement coefficients**

Inter-rater agreement

- Percentage agreement: $\frac{\#\ agreements}{all\ items}$

→ add standard error and confidence intervals

% Agreement A-B: **66%**

% Agreement A-C: **0%**

% Agreement B-C: **33%**

	A	B	C	Agreement AB	Agreement AC	Agreement BC
						
tree – lake	0.5	0.5	0.75	1	0	0
tree – willow	0.75	0.75	1.0	1	0	0
tree – car	0.25	0.0	0.0	0	0	1

Inter-rater agreement

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Percentage agreement: $\frac{\#\ agreements}{all\ items}$

- Inter-rater correlation:

Rater A	Rater B
1	1
2	2
3	3
4	4
5	5

Rater A	Rater C
1	2
2	4
3	6
4	8
5	10

Correlation measures are not very suitable to measure inter-rater agreement!

- Pearson correlation r is 1 in both cases. Output of a system could be scaled down, but these human annotations signal that something is wrong with the task (or with the rater)!

Inter-rater agreement

- Percentage agreement: $\frac{\#\ agreements}{all\ items}$

- Chance-corrected measures: rater C is not applying any knowledge, but is still right sometimes.

Rater A	Rater C
1	1
2	1
3	1
4	1
5	1

Inter-rater agreement

- Percentage agreement: $\frac{\# \, agreements}{all \, items}$
- Chance-corrected measures: rater c is not applying any knowledge, but is still right sometimes.
- Cohen's $\varkappa$ (1960):

$$K = \frac{p0 - pe}{1 - pe}$$

  - p_0: relative observed agreement
  - p_e: probability of chance agreement
  - 5 categories:
    - relative agreement: 0.2
    - chance agreement: 0.2
    - $K = 0$

Rater A	Rater C
1	1
2	1
3	1
4	1
5	1

Inter-rater agreement

- Percentage agreement: $\frac{\#\ agreements}{all\ items}$

- Chance-corrected measures: rater c is not applying any knowledge, but is still right sometimes.
- Cohen's ϰ (1960):

$$\kappa = \frac{p0 - pe}{1 - pe}$$

- Fleiss' ϰ (1982): Generalizes Cohen's ϰ to multiple raters. The basic idea is to consider each pairwise agreement of raters and average over all items i.

Human upper bound

- The inter-rater agreement is often also understood as the human upper bound.
- In many tasks, systems are considered to be perfect once they reach the human upper bound, e.g., machine translation.
- In other tasks, machines should outperform humans, e.g., matrix multiplication.

Summary

Semantic relatedness determines the strength of relationship between two words

Main approaches:

- Path / definition / vector based
- Knowledge based / Distributional

Term-Term & Term-Document Matrix

Evaluation - Pearson/Spearman

- Pearson correlation is highly sensitive to outliers - data inspection is crucial!

Readings

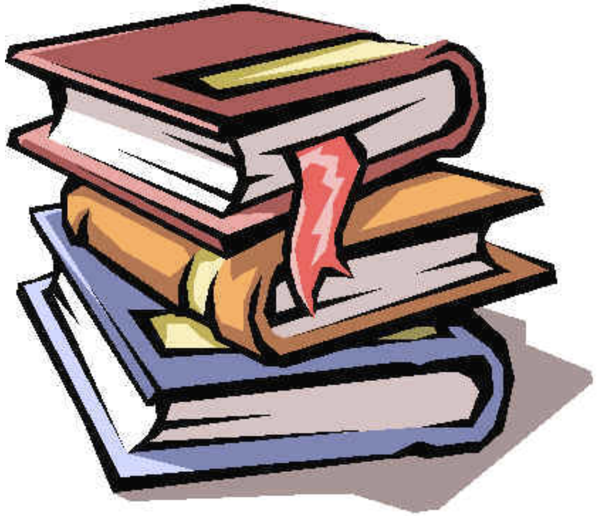
UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Mandatory

- Gärdenfors, Chapter 4:
  - 4.2.1 *Modeling Concepts*, p. 102-105.
  - 4.3 *The Role of Similarity in Concept Formation*, p.109-114.
  - 4.5 *Learning Concepts*, p. 122-126.

Optional

- Gärdenfors, Chapter 3 *Properties*
- Gärdenfors, Chapter 4 *Concepts*



Thank You