

February 21, 2018

Review: Evaluating WordNet-based Measures of Lexical Semantic Relatedness

Alexander Budanitsky*

Graeme Hirst*

Reference: WordNet, <https://dl.acm.org/citation.cfm?id=1168108>

Submitted by: Anoop Mishra



My job for presentation -

- ▶ Reviewing the paper
- ▶ Introducing WordNet
- ▶ Application of WordNet in paper.



Overview

- ▶ Paper description
 - ▶ *Objective*
 - ▶ *Notions used*
- ▶ WordNet Description
 - ▶ *Introduction & History*
 - ▶ *Brief*
- ▶ Paper methodology via WordNet
- ▶ Conclusion

Evaluating WordNet-based Measures of Lexical Semantic relatedness

- ▶ My Objective: To review the paper.
- ▶ This paper belongs to the application of lexical semantic relatedness.

What is Lexical Semantics?

- ▶ i.e. linguistics semantics means study of language meaning.
- ▶ Lexical semantics looks at how the meaning of the lexical units correlates with the structure of the language or syntax. This is referred to as syntax-semantic interface.



Definition: Semantic Relatedness

- ▶ **Semantic relatedness** – any relation between two terms, more general than “**similarity**”.
- ▶ **Difference from semantic similarity** – as this includes only, “**is a**” relations.
- ▶ **Example –**
 - bank-trust company (similar virtue)
 - car-wheel (lexical relationship, *remember – meronyms*)
 - hot-cold (lexical relationship, *remember - antonyms*)
 - pencil-paper (functional relationship)
- ▶ **Computational applications**, required relatedness rather than just similarity.



Applications of measures of relatedness

- ▶ Word sense disambiguation
- ▶ Determining the structure of texts
- ▶ Text summarization & annotation
- ▶ Information extraction and retrieval
- ▶ Automatic Indexing
- ▶ Lexical selection
- ▶ Automatic correction of Word errors and text

Questions framed by author for determining semantic relatedness



Given two measures –

- ▶ Which is good one or bad one?
- ▶ Under what condition, which is better?



▶ **Objective :**

To compare the performance of a number of measures of semantic relatedness for the use of application in NLP & information retrieval.

Notions briefing the objective:

► Notion of relatedness:

Meronymy

Antonymy

Functional Association

Non-classical relations

Example:

Cars & Gasoline

or

Cars & Bicycles



Approach -

Concepts or Words ?





Research Question

Why distributional similarity is not an adequate proxy
for lexical semantic relatedness?

Solutions and Claims

- ▶ All the solutions and claims provided by the authors in the proposal is via **graph modelling**.



- ▶ Authors are using terms of graph like –
*Nodes, links, network, cluster, path length, depth,
taxonomy, regular path, scaling, strongly relations etc.*





1. Proposed Approach

- ▶ Directed graph Approach - Context sensitive [as ref]
- ▶ Roget-Structures Thesauri – Categories (clustering) [as ref]
- ▶ WordNet and Semantic Networks
- ▶ Computing Taxonomic Path length
- ▶ Scaling the network
- ▶ Information based and Integrated Approaches*



Introduction to wordnet

- ▶ **WordNet®** - a large lexical database of English language.
- ▶ **Synsets** - groups English words into set of synonyms, also provide short definitions and usage examples & records number of relations.
- ▶ Hence, Structure of WordNet® can be seen as – dictionary + thesaurus.
- ▶ **Dictionary** – wordbook, collection of words in one or more specific language
- ▶ **Thesaurus** – reference work, list words in grouped to synonyms and sometimes antonyms.



People involved and history

- ▶ **WordNet®** project began in Princeton University at Department of Psychology, and now currently housed at Department of Computer Science.
- ▶ **George A. Miller**, started the project in mid-1980s.
- ▶ It's supported by the grants from the NSF, ARDA, DARPA, DTO, REFLEX and the Tim Gill Foundation.
- ▶ Current members – Christiane Fellbaum, Randee Teng and student collaborators.



Link: <https://wordnet.princeton.edu/>


- Kindly refer above link for **WordNet®** with all the important documents and usage and FAQs.

PRINCETON UNIVERSITY

Search

WordNet

A lexical database for English



What is WordNet?

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.

When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly [cite the source](#). Citation figures are critical to WordNet funding.

About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical

Due to funding and staffing issues, we are no longer able to accept comment and suggestions.

We get numerous questions regarding topics that are addressed on our [FAQ](#) page. If you have a problem or question regarding something you downloaded from the ["Related projects"](#) page, you must contact the developer directly.



WordNet®: Use and applications



► The is used in automatic text analysis and artificial Intelligence applications.

► The database and tools are also freely available to download from the website.



► FrameNet: also a lexical database that refers to Wordnet.

► Lexical Markup framework: standard framework for the construction of lexicons, including WordNet.



► Universal networking project Program: aimed to consolidate lexicosemantic data.



Other languages Used

- ▶ Malayalam WordNet
- ▶ Arabic WordNet
- ▶ Chinese WordNet
- ▶ IndoWordNet – for Indian languages
- ▶ EuroWordNet – for European languages
- ▶ The BalkaNet for six European language
- ▶ Etc.



WordNet Statistics* [1.x]



WordNet is **BIG**.

DESCRIPTION

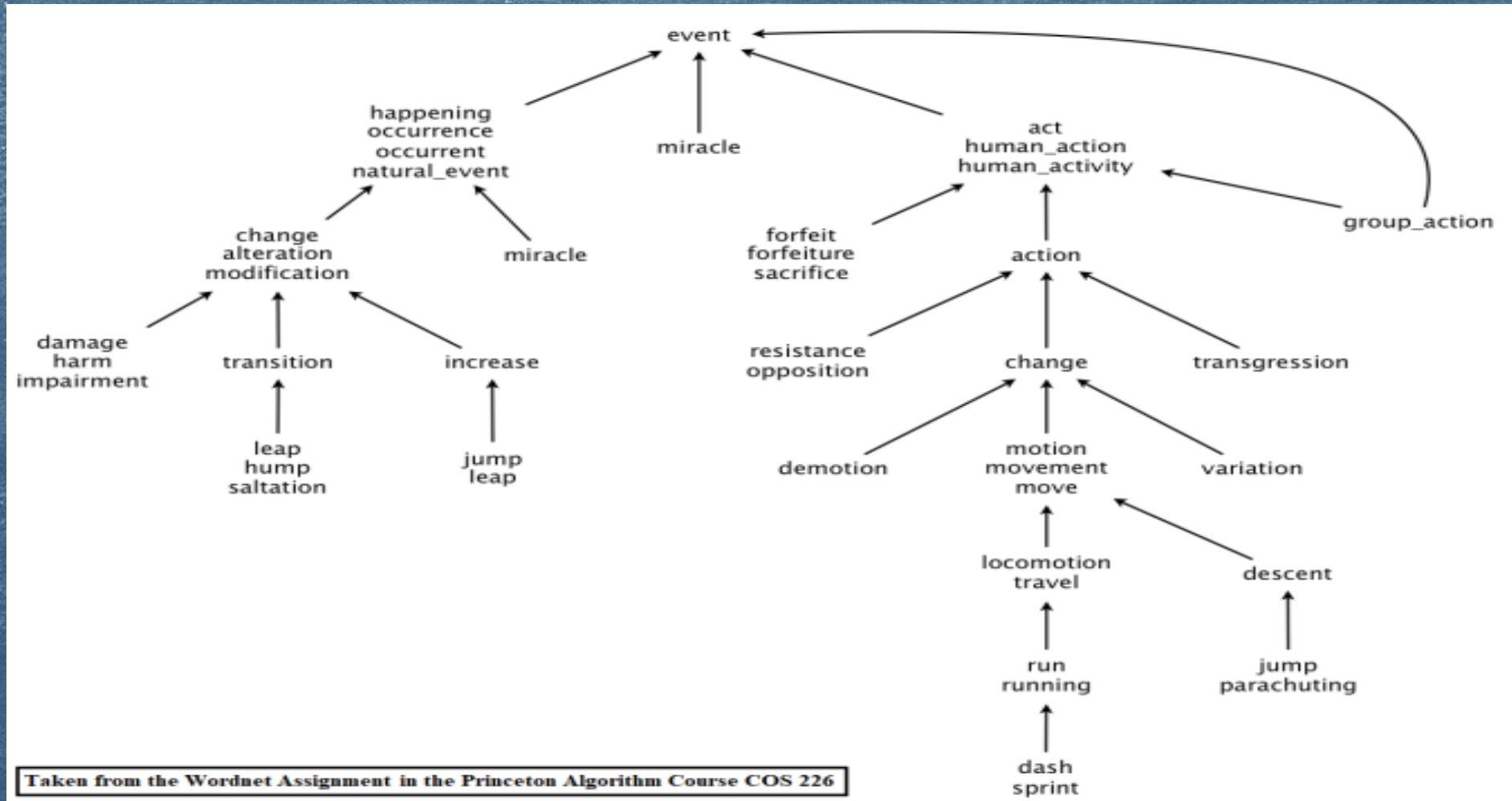
Number of words, synsets, and senses

POS	Unique Synsets		Total
	Strings		Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Polysemy information

POS	Monosemous	Polysemous	Polysemous
	Words and Senses	Words	Senses
Noun	101863	15935	44449
Verb	6277	5252	18770
Adjective	16503	4976	14399
Adverb	3748	733	1832
Totals	128391	26896	79450

What would be the subgraph of WordNet looking like.....



Solution 1: Approaches using WordNet & semantic networks

- ▶ Network of four parts of speech – Noun, Verbs, Adjective & Adverbs. (but not linked, limitation of WordNet 1.x in this paper)
- ▶ **Noun**, is focused as richly developed, accounts for 80% of the relations.

- ▶ **Graph definition:**

Maximum depth = 16 nodes

9 types of relations – hyponymy & inverse, hypernymy, 6-meronyms

$\text{len}(C_i, C_j)$ – length of shortest path from synset C_i to C_j .

$\text{depth}(C_i) = \text{len}(\text{root}, C_i)$ – length of path from the global root.

$\text{Iso}(C_1, C_2)$ – most specific common ancestor of C_1, C_2



Formula for semantic relatedness in Solution 1

$$\text{rel}(w_1, w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} [\text{rel}(c_1, c_2)]$$

- ▶ $\text{rel}(C_1, C_2)$ – semantic relatedness b/w C_1 & C_2
- ▶ $\text{rel}(w_1, w_2)$ – semantic relatedness b/w w_1 & w_2
- ▶ $S(w_i)$ – senses of the words
- ▶ i.e. the relatedness of two words is equal to that of the most-related pair of concepts that they denote.



Solution 2: Computing Taxonomic Path Length

- ▶ Regarded as, “simple way to compute semantic relatedness in a taxonomy such as WordNet ”.
- ▶ “The shorter the path from one node to another, the more similar they are” (Resnik 1995).
- ▶ Relation - Medium, Strong or Regular (if exists an allowable path)
- ▶ allowable path - if it contains no more than five links - “the longer the path and the more changes of direction, the lower the weight”.
- ▶ allowable path may include more than one link and that the directions of links on the same path may vary (among upward (hypernymy and meronymy), downward (hyponymy & holonymy) and horizontal (antonymy)).



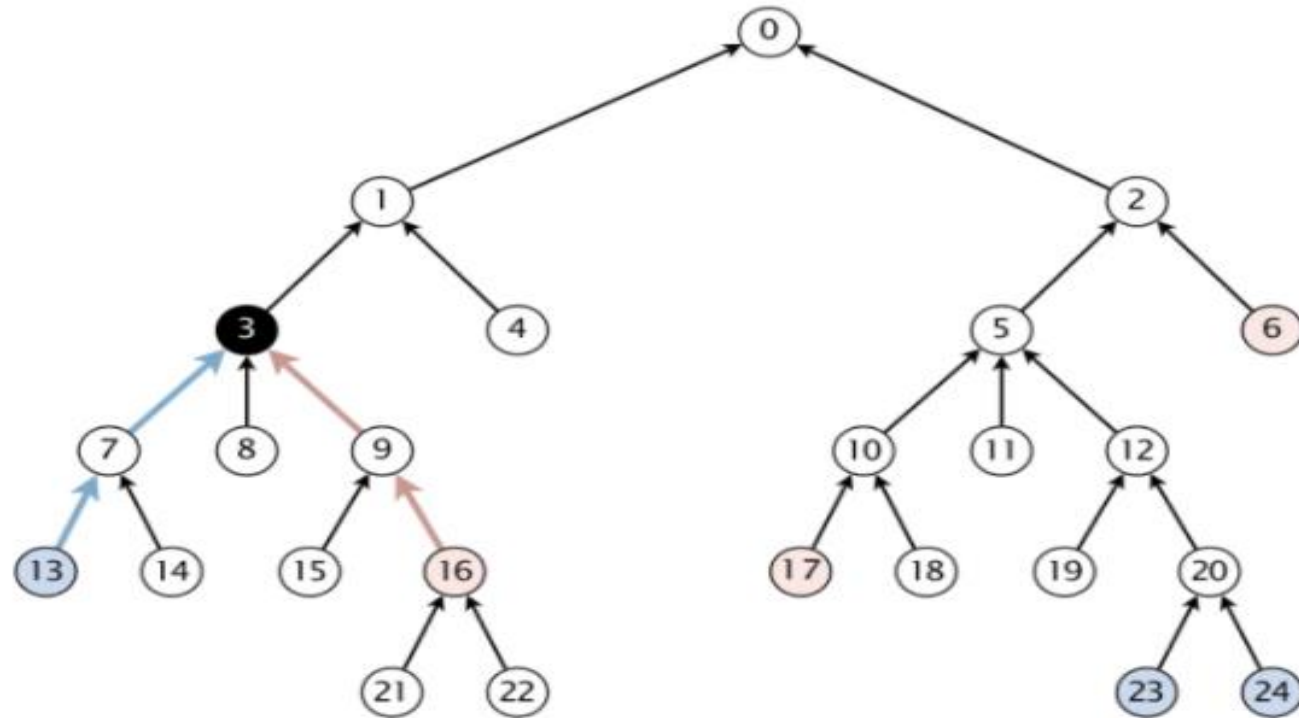
Formula for semantic relatedness in Solution 2

- ▶ As per Hirst and St-Onge's approach formula for two WordNet concepts $c_1 \neq c_2$

$$\text{rel}_{\text{HS}}(c_1, c_2) = C - \text{len}(c_1, c_2) - k \times \text{turns}(c_1, c_2)$$

- ▶ C & k are constants
- ▶ (in practice $C = 8$ and $k = 1$), and $\text{turns}(c_1, c_2)$ is the number of times the path between c_1 and c_2 changes direction.

Ancestral path concept (depth) -



$A = \{ 13, 23, 24 \}$, $B = \{ 6, 16, 17 \}$
ancestral path: 13-7-3-1-0-2-6
ancestral path: 23-20-12-5-10-17
ancestral path: 23-20-12-5-2-6

shortest ancestral path: 13-7-3-9-16
associated length: 4
shortest common ancestor: 3

Solution 3: Leacock and Chodorow's Normalized Path Length

- ▶ Leacock and Chodorow (1998) proposed the following formula for computing the scaled semantic similarity between concepts c_1 and c_2 in WordNet:



$$\text{sim}_{\text{LC}}(c_1, c_2) = -\log \frac{\text{len}(c_1, c_2)}{2 \times \max_{c \in \text{WordNet}} \text{depth}(c)}$$

- ▶ The denominator includes the maximum depth of the hierarchy.



Solution 4: Resnik's Information-based Approach

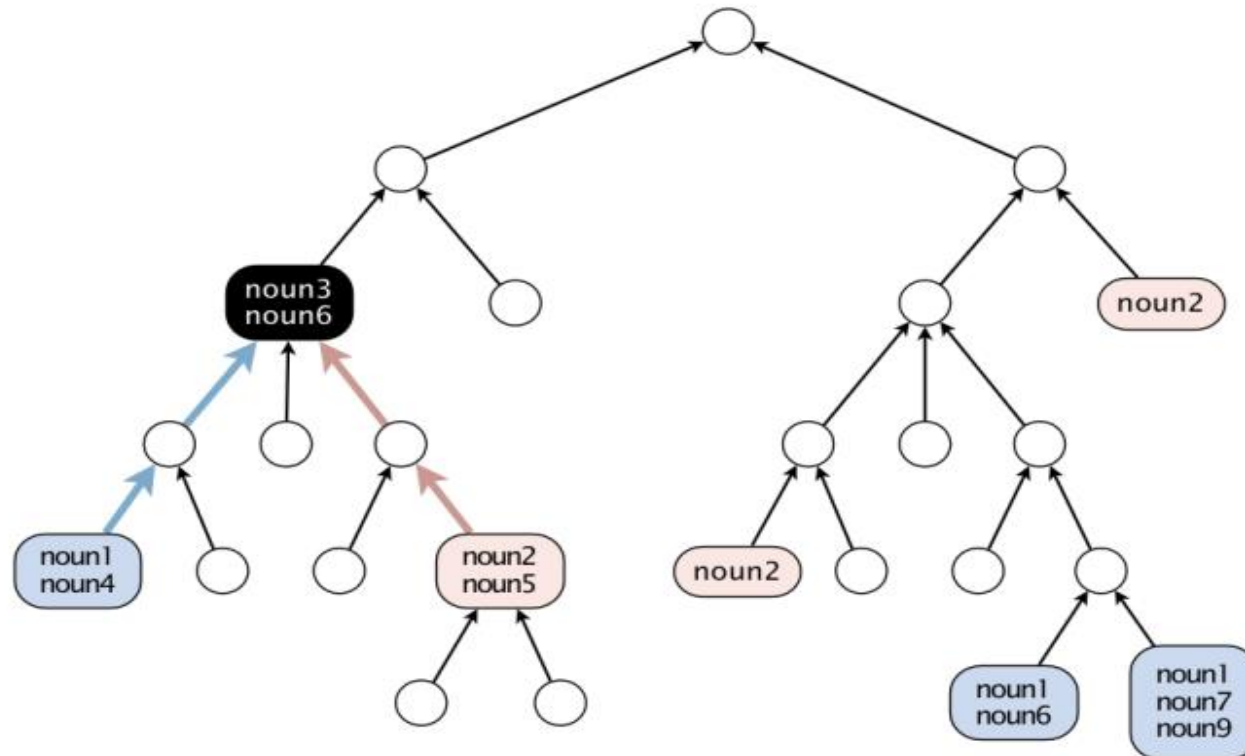
- ▶ Similarity between two concepts is “the extent to which they share information in common”,
- ▶ Determined by inspecting the relative position of the most-specific concept that subsumes them both.
- ▶ semantic similarity of a pair of concepts $c1$ and $c2$,

$$p(c) = \frac{\sum_{w \in W(c)} \text{count}(w)}{N}$$

- ▶ $W(c)$ is the set of words (nouns) in the corpus whose senses are subsumed by concept c , and N is the total number of word (noun) tokens in the corpus that are also present in WordNet.



In terms of noun relations:



$\text{distance}(\text{noun1}, \text{noun2}) = 4$

$\text{sca}(\text{noun1}, \text{noun2}) = \{\text{noun3}, \text{noun6}\}$

Taken from the Wordnet Assignment in the Princeton Algorithm Course COS 226

Solution 5: Jiang and Conrath's Combined Approach

- ▶ Reacting to the disadvantages of Resnik's method
- ▶ to synthesize edge- and node based techniques by restoring network edges to their dominant role in similarity computations, and using corpus statistics as a secondary, corrective fact.
- ▶ the semantic distance between concepts c_1 and c_2 is the sum of the distances along the shortest path that connects the nodes:

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



Solution 6: Lin's Universal Similarity Measure

- ▶ define a measure of similarity that would be both universal and theoretically justified.
- ▶ **Similarity Theorem:** 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.
- ▶ His measure of similarity between two concepts in a taxonomy is a corollary of this theorem –

$$\text{sim}_L(c_1, c_2) = \frac{2 \times \log p(\text{lso}(c_1, c_2))}{\log p(c_1) + \log p(c_2)}$$

- ▶ where the probabilities $p(c)$ are determined in a manner analogous to Resnik's $p(c)$ [Solution 3]



Snippets for graph modelling using Wordnet

Ref-

<https://github.com/GradySimon/SemanticRelatedness/blob/master/relatednesstest.py>

```
dog_entities = ["dog", "cat", "horse", "saddle", "rider", "mouse", "cheese",
               "churning", "milk", "cow", "human", "race", "gambling"]
dog_connections = [("dog", "cat", 50),
                  ("dog", "horse", 10),
                  ("horse", "saddle", 60),
                  ("horse", "rider", 30),
                  ("rider", "saddle", 40),
                  ("horse", "race", 30),
                  ("rider", "race", 35),
                  ("dog", "race", 20),
                  ("cat", "mouse", 50),
                  ("race", "gambling", 40),
                  ("mouse", "cheese", 50),
                  ("cheese", "milk", 60),
                  ("milk", "cow", 60),
                  ("cheese", "cow", 30),
                  ("rider", "human", 50),
                  ("milk", "churning", 20),
                  ("human", "gambling", 30)]

def get_test_network1():
    """ Returns a sample Semantic_Network.
    """
    network = Semantic_Network()
    network.add_entities(dog_entities)
    network.add_connections(dog_connections)
    return network

def draw_network(network, location=None):
    """ Displays in an interactive window a visualization of the graph underlying the specified Semantic_Network.
    """
    graph = network.graph
    draw_graph(graph, location)

def draw_graph(graph, location=None):
```




Evaluation Methods

- ▶ Authors reasoned about and evaluate computational measures of semantic
- ▶ relatedness via three kind of approaches –
- ▶ 1st - Mathematically, parameter-projections are smooth function
- ▶ 2nd - comparison with human judgments (best assessment of “goodness” of measure)
- ▶ 3rd - measures with respect to their performance in framework of a particular appⁿ

-
- ▶ This paper, used the **second** and the **third** methods to compare several different measures and use WordNet as their knowledge resource.
 - ▶ **Another evaluation method used- Detect Malapropisms**
Malapropisms – measures are compared through performance of the application that uses, the detection and correction of real-word spelling errors in open-class words.



Valuation definitions and Mean Results

- Data from – Rubenstein and Goodenough (1965) (R&G)
Miller and Charles (1991) (M&C)

Measure	M&C	R&G
Hirst and St-Onge, rel_{HS}	.744	.786
Jiang and Conrath, $dist_{JC}$.850	.781
Leacock and Chodorow, sim_{LC}	.816	.838
Lin, sim_L	.829	.819
Resnik, sim_R	.774	.779

Graph Chart used: [Reference – Brown Corpus]

Table 1

Human and computer ratings of the Rubenstein–Goodenough set of word pairs (*part 1 of 2*).

#	Pair		Humans	rel _{HS}	dist _{JC}	sim _{LC}	sim _L	sim _R
1	cord	smile	0.02	0	19.6	1.38	0.09	1.17
2	rooster	voyage	0.04	0	26.9	0.91	0.00	0.00
3	noon	string	0.04	0	22.6	1.50	0.00	0.00
4	fruit	furnace	0.05	0	18.5	2.28	0.14	1.85
5	autograph	shore	0.06	0	22.7	1.38	0.00	0.00
6	automobile	wizard	0.11	0	17.8	1.50	0.09	0.97
7	mound	stove	0.14	0	17.2	2.28	0.22	2.90
8	grin	implement	0.18	0	16.6	1.28	0.00	0.00
9	asylum	fruit	0.19	0	19.5	2.28	0.14	1.85
10	asylum	monk	0.39	0	25.6	1.62	0.07	0.97
11	graveyard	madhouse	0.42	0	29.7	1.18	0.00	0.00
12	glass	magician	0.44	0	22.8	1.91	0.07	0.97
13	boy	rooster	0.44	0	17.8	1.50	0.21	2.38
14	cushion	jewel	0.45	0	22.9	2.28	0.13	1.85
15	monk	slave	0.57	94	18.9	2.76	0.21	2.53
16	asylum	cemetery	0.79	0	28.1	1.50	0.00	0.00
17	coast	forest	0.85	0	20.2	2.28	0.12	1.50
18	grin	lad	0.88	0	20.8	1.28	0.00	0.00
19	shore	woodland	0.90	93	19.3	2.50	0.13	1.50
20	monk	oracle	0.91	0	22.7	2.08	0.18	2.53
21	boy	sage	0.96	93	19.9	2.50	0.20	2.53
22	automobile	cushion	0.97	98	15.0	2.08	0.27	2.90
23	mound	shore	0.97	91	12.4	2.76	0.49	6.19
24	lad	wizard	0.99	94	16.5	2.76	0.23	2.53
25	forest	graveyard	1.00	0	24.5	1.76	0.00	0.00
26	food	rooster	1.09	0	17.4	1.38	0.10	0.97
27	cemetery	woodland	1.18	0	25.0	1.76	0.00	0.00
28	shore	voyage	1.22	0	23.7	1.38	0.00	0.00
29	bird	woodland	1.24	0	18.1	2.08	0.13	1.50
30	coast	hill	1.26	94	10.8	2.76	0.53	6.19



Results discussion -

- ▶ **1st** - while semantic relatedness is inherently a relation on concepts.
- ▶ **2nd** - semantic relatedness is symmetric, distributional similarity is a potentially asymmetrical relationship.
- ▶ **3rd** - lexical semantic relatedness depends on pre-defined lexicographic or other knowledge resource, whereas distributional similarity is relative to a corpus.



Limitations

- ▶ In practice, tiny amount of data available is quite inadequate.
- ▶ Constructing a large-enough set of pairs and obtaining human judgments on them would be a very large task.
- ▶ Interested in is the relationship between the concepts for which the words are merely surrogates.
- ▶ The human judgments that we need are of the relatedness of word-senses, not words. But the experimental situation would need to set up contexts that bias the sense selection for each **target word**.



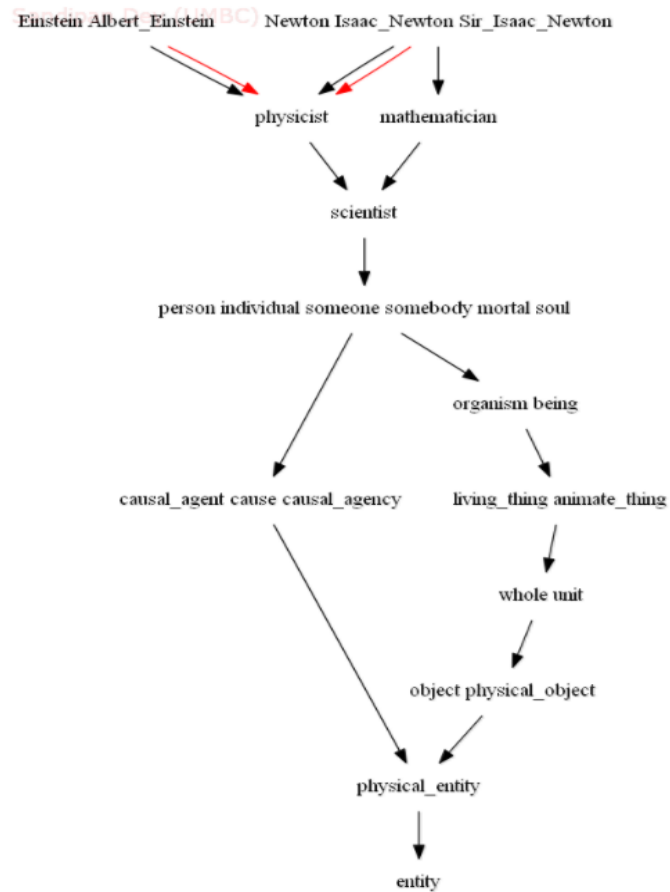
Conclusion of the paper

- ▶ Lexical semantic relatedness is more general idea of **relatedness** not just the **similarity**.
- ▶ These relationships include not just hyponymy and the non-hyponymy relationships in WordNet such as meronymy but also **associative** and **ad-hoc** relationships. These can include just about any kind of **functional relation** or **frequent association in the world**.



Relate the conclusion.....

Einstein and Newton (distance 2 in the Wordnet Digraph, with SCA as *physicist* as shown below)





What I learnt -

- ▶ Hence, lexical semantic relatedness is sometimes constructed in **context** and cannot always be determined purely from an a priori lexical resource such as WordNet.

Name	Example
IS-USED-TO	<i>bed-sleep</i>
WORKS-IN	<i>judge-court</i>
LIVES-IN	<i>camel-desert</i>
IS-THE-OUTSIDE-OF	<i>husk-corn</i>

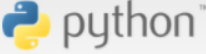
References

- ▶ Evaluating WordNet-based Measures of Lexical Semantic Relatedness, <https://dl.acm.org/citation.cfm?id=1168108>
- ▶ WordNet, <https://wordnet.princeton.edu/>
- ▶ <http://www.nltk.org/book/>
- ▶ <http://www.cs.princeton.edu/courses/archive/spring18/cos226/lectures.php>

Thank You!

Questions ?

Extra

[Check out the all new PyPI! \(More information here\)](#)

» Package Index > sematch > 1.0.4

PACKAGE INDEX >>

[Browse packages](#)

[List trove classifiers](#)

[RSS \(latest 40 updates\)](#)

[RSS \(newest 40 packages\)](#)

[Terms of Service](#)

[PyPI Tutorial](#)

[PyPI Security](#)

[PyPI Support](#)

[PyPI Bug Reports](#)

[PyPI Discussion](#)

[PyPI Developer Info](#)

ABOUT >>

NEWS >>

DOCUMENTATION >>

DOWNLOAD >>

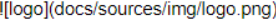
COMMUNITY >>

FOUNDATION >>

CORE DEVELOPMENT >>

sematch 1.0.4

Semantic similarity framework for knowledge graphs

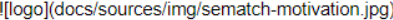


[Download sematch-1.0.4.tar.gz](#)

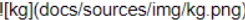
Introduction

Se-match is an integrated framework for the development, evaluation, and application of semantic similarity for Knowledge Graphs (KGs). It is easy to use Se-match to compute semantic similarity scores of concepts, words and entities. Se-match focuses on specific knowledge-based semantic similarity metrics that rely on structural knowledge in taxonomy (e.g. depth, path length, least common subsumer), and statistical information contents (corpus-IC and graph-IC). Knowledge-based approaches differ from their counterpart corpus-based approaches relying on co-occurrence (e.g. Pointwise Mutual Information) or distributional similarity (Latent Semantic Analysis, Word2Vec, GLOVE and etc). Knowledge-based approaches are usually used for structural KGs, while corpus-based approaches are normally applied in textual corpora.

In text analysis applications, a common pipeline is adopted in using semantic similarity from concept level, to word and sentence level. For example, word similarity is first computed based on similarity scores of WordNet concepts, and sentence similarity is computed by composing word similarity scores. Finally, document similarity could be computed by identifying important sentences, e.g. TextRank.



KG based applications also meet similar pipeline in using semantic similarity, from concept similarity (e.g. "http://dbpedia.org/class/yago/Actor109765278") to entity similarity (e.g. "http://dbpedia.org/resource/Madrid"). Furthermore, in computing document similarity, entities are extracted and document similarity is computed by composing entity similarity scores.



In KGs, concepts usually denote ontology classes while entities refer to ontology instances. Moreover, those concepts are usually constructed into hierarchical taxonomies, such as DBpedia ontology class, thus quantifying concept similarity in KG relies on similar semantic information (e.g. path length, depth, least common subsumer, information content) and semantic similarity metrics (e.g. Path, Wu & Palmer, Li, Resnik, Lin, Jiang & Conrad and WPath). In consequence, Se-match provides an integrated framework to develop and evaluate semantic similarity metrics for

Not Logged In

[Log in](#)

[Register](#)

[Lost Login?](#)

[Log in with OpenID](#)

[Log in with Google](#)

Status

[TlSv1.0/TlSv1.1 rolling](#)

[brownouts is currently in](#)

[process](#)

42