



Limits of Lexical Semantic Relatedness with Ontology-based Conceptual Vectors

Lian-Tze Lim, Didier Schwab

► To cite this version:

Lian-Tze Lim, Didier Schwab. Limits of Lexical Semantic Relatedness with Ontology-based Conceptual Vectors. 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008), Jun 2008, Barcelone, Spain. hal-03319019

HAL Id: hal-03319019

<https://hal.science/hal-03319019>

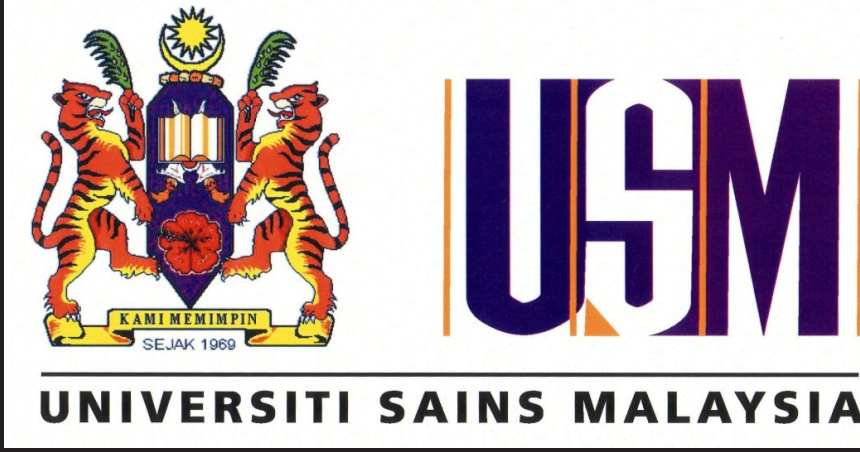
Submitted on 11 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Limits of Lexical Semantic Relatedness with Ontology-based Conceptual Vectors

Lian-Tze Lim (Universiti Sains Malaysia)
Didier Schwab (Université Pierre Mendès France)



Introduction

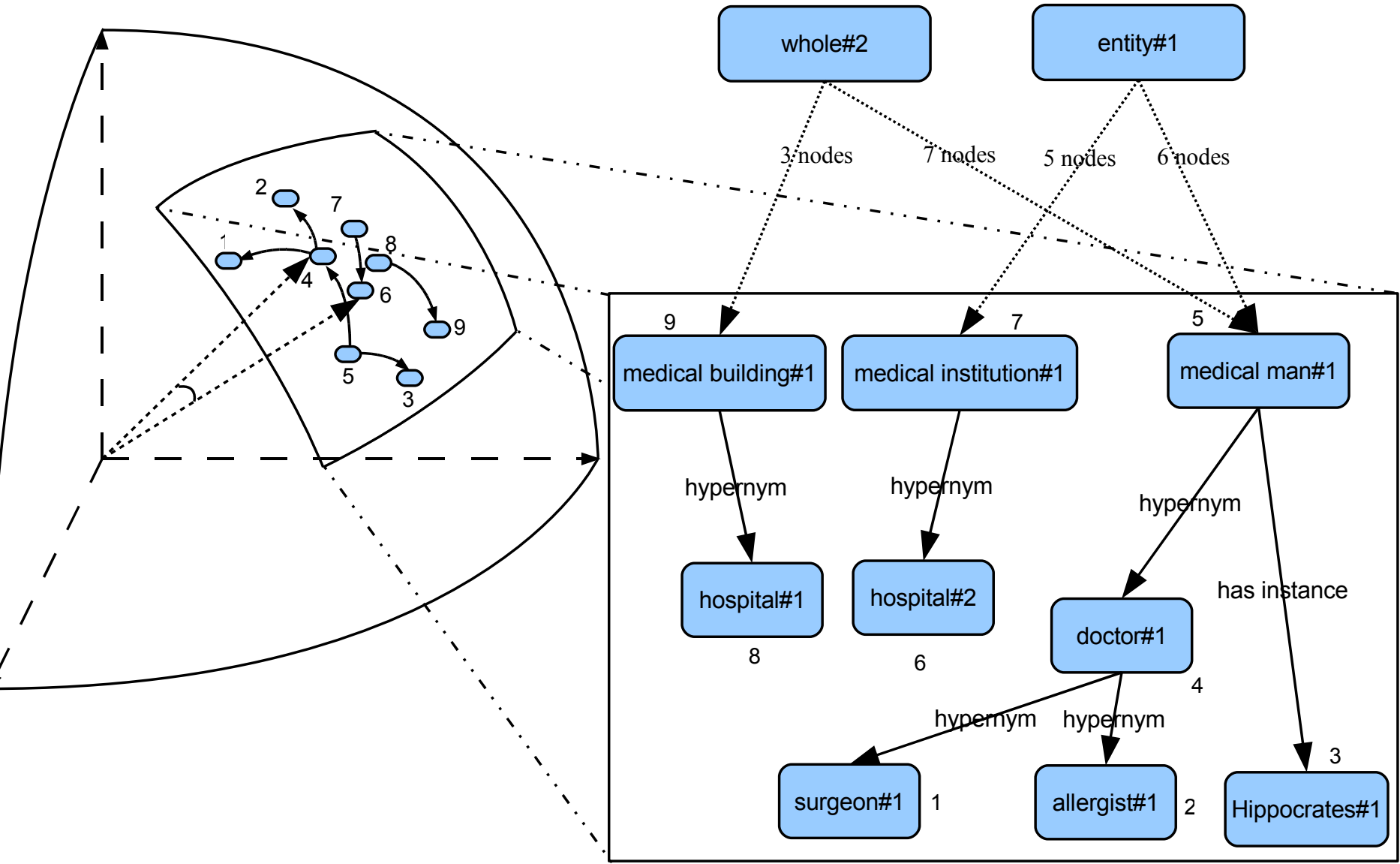
Miller-Charles dataset (1991)

- ▶ Human judgements of similarity between pairs of English words
- ▶ Machine-computed scores can be compared against dataset

We aim to

- ▶ Represent thematic aspects of text (incl. word senses) with **Conceptual Vectors**
- ▶ Define semantic relatedness based on conceptual vectors
- ▶ Study behavior of CVs constructed based on ontology
- ▶ ... by comparing against Miller-Charles dataset

Complementing WordNet with Non-discrete Navigation



- ▶ WordNet: explicit semantic relations between senses
- ▶ Introduce **non-discrete** navigation with idea of **neighbourhood**
- ▶ To solve (in part) Tennis Problem of WordNet
- ▶ **CVs for WordNet senses** projects them onto **hyperspace**

Conceptual Vectors (CV)

Principle and Thematic Distance

- ▶ Inspired by **componential semantics**
 - ▶ Formalism projecting semantic components into vectorial space
 - ▶ CV elements correspond to concepts indirectly
 - ▶ Overlap of shared ideas between word senses X & Y :
Thematic proximity: $Sim(X, Y) = \cos(\overline{X}, \overline{Y}) = \frac{X \cdot Y}{\|X\| \times \|Y\|}$
Angular distance: $D_A = \arccos(Sim(A, B))$

Operations on Vectors

- ▶ **Normalised Vectorial Sum** ($\oplus$): averages operand vectors
- ▶ **Vectorial Term-to-Term Product** ($\odot$): highlights common ideas
- ▶ **Weak Contextualisation**: emphasizes features shared by two terms, accentuated by each other

$$\gamma(X, Y) = X \oplus (X \odot Y)$$

- ▶ **Synonymy**: tests thematic closeness of two meanings X and Y , each enhanced with what it has in common with a third (C)

$$Syn_R(X, Y, C) = D_A(\gamma(X, C), \gamma(Y, C))$$

- ▶ **Partial Synonymy**: Syn_R where the context is the sum of contextualisation of X and Y of their means

$$Syn_P(X, Y) = Syn_R(X, Y, \gamma(X, X \oplus Y) \oplus \gamma(Y, X \oplus Y))$$

Construction

- ▶ Can be computed from definitions from different sources (dictionaries, synonym lists, hand-crafted indices, ...)
- ▶ Fabricates new CVs from existing CVs
→ requires bootstrap kernel of pre-computed CVs
- ▶ Emergence from randomised vectors (Lafourcade 2006)
 - ▶ dimension of the vector space can be chosen freely
 - ▶ more constant lexical density in vector space
 - ▶ no prior sense-to-concept mapping required
 - ▶ **requires much more iterations, time and computing resources**
- ▶ We try with **ontology-based CVs** first

Constructing Ontology-based CV for WordNet Senses

Vector Construction for Ontology Classes

- ▶ 1992 elements (each correspond to class in SUMO/MILO (not all were used)) in CV
- ▶ For each ontology class $\mathcal{C}$, compute $V(\mathcal{C}) = (v_1(\mathcal{C}), v_2(\mathcal{C}), v_3(\mathcal{C}), \dots)$
 - ▶ Initialise $V^0(\mathcal{C})$ with $v_i^0(\mathcal{C}) = 1$ if v_i corresponds to $\mathcal{C}$, 0 otherwise
 - ▶ Then $V(\mathcal{C})$ with $v_i(\mathcal{C}) = v_i^0(\mathcal{C}) + \sum_{j=1}^{dim(V)} \frac{v_j^0(\mathcal{C})}{2^{dist(\mathcal{C}, \mathcal{C}_j)}}$

Constructing Ontology-based CV for WordNet Senses (cont.)

Kernel Vectors for WordNet Synsets

- ▶ For each WN sense s , initialise $V^0(s) = V(\mathcal{C})$; $\mathcal{C}$ is the SUMO/MILO class assigned to s (Niles & Pease, 2003)

Learning CVs for WN Senses

- ▶ Iteratively compute $V(s)$ as in (Schwab *et al*, 2007) but using $V^0(s)$ based on eXtended WordNet

Comparison with Miller-Charles (M&C) Set

- ▶ We define two proximity measures based on CV:

$$\text{prox}_{cv}(A, B) = 1 - \left(D_A(A, B) \div \frac{\pi}{2}\right)$$

$$\text{prox}_{syn}(A, B) = 1 - \left(Syn_P(A, B) \div \frac{\pi}{2}\right)$$

- ▶ Take highest prox_{cv} and prox_{syn} values of WN noun senses

Word Pair	M&C	prox_{cv}	prox_{syn}	Corr. with M&C	
				prox_{cv}	prox_{syn}
automobile car	0.98	1.00	1.00	1.000	1.000
cord smile	0.03	0.48	0.57	1.000	1.000
glass magician	0.03	0.47	0.57	1.000	1.000
gem jewel	0.96	1.00	1.00	1.000	1.000
rooster voyage	0.02	0.44	0.53	0.999	0.998
magician wizard	0.88	0.90	0.92	0.997	0.997
bird crane	0.74	0.82	0.86	0.996	0.996
crane implement	0.42	0.60	0.68	0.989	0.991
noon string	0.02	0.38	0.47	0.988	0.988
bird cock	0.76	0.78	0.83	0.983	0.985
coast shore	0.93	0.86	0.89	0.980	0.982
journey voyage	0.96	0.85	0.88	0.974	0.976
midday noon	0.86	1.00	1.00	0.965	0.968
furnace stove	0.78	0.70	0.77	0.950	0.956
implement tool	0.74	0.67	0.74	0.936	0.944
brother lad	0.42	0.47	0.55	0.925	0.930
food rooster	0.22	0.34	0.43	0.920	0.921
lad wizard	0.11	0.20	0.30	0.915	0.911
asylum madhouse	0.90	0.70	0.76	0.901	0.897
food fruit	0.77	0.60	0.69	0.885	0.885
boy lad	0.94	0.62	0.70	0.856	0.860
monk slave	0.14	0.62	0.69	0.837	0.839
car journey	0.29	0.71	0.77	0.818	0.818
monk oracle	0.28	0.71	0.77	0.800	0.798
coast hill	0.22	0.71	0.77	0.777	0.774
forest graveyard	0.21	0.81	0.85	0.735	0.731
coast forest	0.11	0.70	0.77	0.711	0.704
brother monk	0.71	0.34	0.43	0.644	0.634
Corr. with M&C		0.644	0.634		

Discussion

Results are not too impressive...

Suitability of WN-SUMO/MILO mappings

- ▶ *brother-monk* scored too low because mapped to very different classes (HUMAN and RELIGIOUSORGANISATION)
- ▶ *coast-hill*, *forest-graveyard* and *coast-forest* considered dissimilar by humans, but all LANDAREA in mapping
- ▶ (Take these away and correlations with M&C rise to **0.800** and **0.798**)

Suitability of the M&C Set

- ▶ *cars* and *gasoline* are semantically **related**;
cars and *bicycles* are semantically **similar** (Resnik, 1995)
- ▶ *car-journey*: low M&C score; high prox_{cv} and prox_{syn} scores
- ▶ M&C assesses **semantic similarity**
- ▶ CVs (via prox_{cv} and prox_{syn}) assesses **semantic relatedness**

Conclusion

- ▶ CVs can model the ideas conveyed by lexical meanings
- ▶ Can be constructed based on ontological sources
- ▶ Can be used to measure lexical semantic relatedness
- ▶ Disadvantage of ontology-based CVs:
 - ▶ non-standard density of the hierarchy
 - ▶ different philosophies in mapping lexical senses to ontology classes

Future Work

- ▶ Effects of hierarchy-free CVs i.e. construction by emergence
- ▶ Collect human ratings on lexical semantic *relatedness* as benchmark