Domain-Specific Semantic Relatedness from Wikipedia Structure: A Case Study in Biomedical Text

Armin Sajadi, Evangelos Milios, Vlado Kešelj, and Jeannette Janssen
{sajadi,eem,vlado}@cs.dal.ca, janssen@mathstat.dal.ca

April 18, 2015

1. Domain Specific Relatedness

- Calculating relatedness between two domain-specific concepts.
- The relation can be taxonomic relation (i.e., *is-a*) or any non taxonomic relation such as *is-treated-by* in the biomedical domain.
- Measuring relatedness benefits NLP applications.

2. Contributions

- Comparing Wikipedia in the biomedical domain with both (1) Ontology-based methods and (2) distributional methods.
- Evaluating a group of graph-based similarity methods on Wikipedia.
- Proposing a new relatedness method using Wikipedia graph structure.

3. Motivations

3.1. Why Wikipedia?

- Domain-specific semantic relatedness relies on either ontologies or specialized corpora.
- Ontologies are labor-intensive and do not exist for most domains.
- Distributional methods need sufficiently large domain specific corpora. Building such corpora is not trivial.

3.2. Why Biomedical Domain?

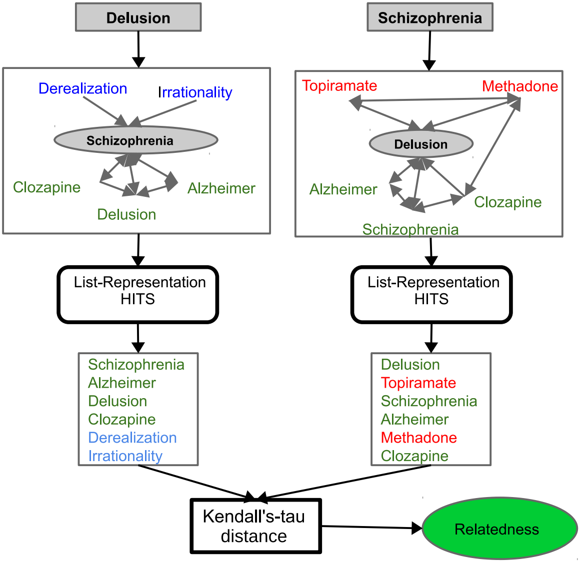
- The availability of high-quality ontologies (MeSH, SNOMED-CT, etc.).
- A rich literature for extracting semantic relatedness.
- The availability of reliable datasets.

4. Basic Idea

Given two concepts:

- Extract neighborhood graph for each concept in the Wikipedia graph.
- Transform the graph to a list using HITS algorithm.
- Calculate Kendall's tau distance between the two lists.

5. Relatedness Calculation



6. Formulation

6.1. HITS Ranking Algorithm: Originally proposed to rank web pages

- Input: A graph with adjacency matrix M
- Output: two scoring functions on vertices: Authorities and Hubs
- Idea: Mutual Reinforcement

$$\begin{aligned} \textit{Hub-scores} &\leftarrow \text{Principal Eigen-vector of } M^T M \\ \textit{Auth-scores} &\leftarrow \text{Principal Eigen-vector of } M M^T \end{aligned}$$

6.2. Kendall's tau Distance:
Counts the number of pairwise disagreements between two given lists σ_1 and σ_2 :

$$K(\sigma_1, \sigma_2) = \frac{2}{n(n-1)} \sum_{\{i,j\} \in \mathcal{P}} \bar{K}_{i,j}(\sigma_1, \sigma_2)$$

where

- $\mathcal{P}$ is the set of unordered pairs of distinct elements of the lists
- $K_{i,j}(\sigma_1, \sigma_2)$ is 0 if i and j are in the same order in both of the lists, otherwise it is 1

6.3. HITS-Sim Score:

$$\begin{aligned} \mathit{HITS\text{-}sim}(a,b) &= \lambda \times \mathit{HITS\text{-}sim}_{hub}(a,b) \\ &+ (1-\lambda) \times \mathit{HITS\text{-}sim}_{aut}(a,b) \\ & \quad \lambda \in [0,1] \text{ (we use 0.5)} \end{aligned}$$

Table 1. Comparison with Ontology-based methods. $\mathbf{o_1}$: *sct-umls*; $\mathbf{o_2}$: *mesh-umls*; $\mathbf{o_3}$: *umls*

Method	**Pedersen. N=29** [2]			**Mayo N=101** [4]			**UMN sim. N=566** [3]			**UMN rel N=587** [3]		
	$\mathbf{o_1}$	$\mathbf{o_2}$	$\mathbf{o_3}$	$\mathbf{o_1}$	$\mathbf{o_2}$	$\mathbf{o_3}$	$\mathbf{o_1}$	$\mathbf{o_2}$	$\mathbf{o_3}$	$\mathbf{o_1}$	$\mathbf{o_2}$	$\mathbf{o_3}$
LCH	.44	.42	.61	.03	.26	.3	.23	.25	.4	.17	.34	.34
IIC-LCH	.38	.43	.7	.3	.25	.44	.36	.29	.46	.3	.35	.39
PPR	.63	.31	.69	.17	.05	.46	.23	.18	.41	.17	.18	.33
hits-sim	**.71**			***.52**			†**.58**			†**.51**		

Table 2. Comparison with distributional methods

Method	Resources	Pedersen	Mayo	UMN sim.	UMN rel.
Vector	Mayo Corpus* +UMLS	.76	†.02	†.02	†-.13
Tensor	OHSUMED+UMLS	.76			
Word2Vec	OHSUMED	†.34	†.26	†.36	†.29
Word2Vec	OHSUMED+UMLS	**.80**	**.63**	†.39	†.39
hits-sim	Wikipedia	.71	.52	**.58**	**.51**

* Mayo Corpus of Clinical Notes.

Table 3. Comparison between Wikipedia based methods

Method	MC [1]	WordSim353 [5]	Ped. Phys.	Ped. Coders	Ped. All	Mayo	UMN Sim.	UMN Rel.
ESA	.73	**.75**						
CPRel	.83	.64						
WLM†	.86	.67	.63	.69	.67	.49	**.58**	.49
Co-Citation†	.86	.67	.62	.68	.66	.47	.57	.49
Coupling†	**.90**	*.65	.61	.66	.64	*.44	*.49	*.4
Amsler†	.86	.68	.58	.66	.64	*.45	*.53	*.43
SimRank†	.79	*.51	*.56	*.55	*.55	*.39	*.45	*.39
EHITS-sim†	.84	*.62	.6	.67	.64	*.46	*.54	*.45
HITS-sim	.88	.70	**.67**	**.72**	**.71**	**.52**	**.58**	**.51**

Table 4. The Effect of Metrics: Kendall's tau (τ), Pearson (r) and *cosine* distance (*cos*)

	Pedersen			MayoSRS			UMN Rel.			UMN Sim.		
	τ	r	*cos*	τ	r	*cos*	τ	r	*cos*	τ	r	*cos*
ρ	**.71**	.57	.64	**.52**	.42	**.52**	**.58**	.35	.55	**.51**	.36	.49

8. Conclusion

- Distributional and ontology-based methods are competitive, and a hybrid of them improves the results.
- Wikipedia is comparable with the available specialized resources and often significantly improves upon them.
- Our new proposed graph-based relatedness computing approach based on the HITS algorithm achieves the best correlations with human judgement.

8. References I

Miller et al. (1991).
Contextual correlates of semantic similarity.
Language and Cognitive Processes, 6(1):1–28.

Pedersen et al. (2007).
Measures of semantic similarity and relatedness in the biomedical domain.
Biomedical Informatics, 40(3):288 - 299.

Pakhomov et al. (2010).
Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study.
AMIA Annu Symp Proc, 2010:572–576.

Pakhomov et al. (2011).
Towards a framework for developing semantic relatedness reference standards.
Biomedical Informatics, 44(2):251–265.

Finkelstein (2001).
Placing search in context: the concept revisited.
Conference on World Wide Web, 406–414.