Computational Lexical Semantics: Methods and Applications

Alexander Panchenko
Language Technology Group, TU Darmstadt, Germany
panchenko@lt.informatik.tu-darmstadt.de

November 12, 2015

TECHNISCHE UNIVERSITÄT DARMSTADT

Plan

Plan

Similarity Measures

- **Similarity measure** is a numerical measure of the degree the two objects are alike
- **Dissimilarity measure** is a numerical measure of the degree to which the two objects are different.
- Both similarity and dissimilarity scores are scalars in range $[0; 1]$ or $[0; \infty]$.
- Two similar objects i and j will have a high similarity score s_{ij} and a low dissimilarity score d_{ij}.
- Similarity to dissimilarity and vice versa:
  - if $d_{ij} \in [0; 1]$, then $s_{ij} = 1 - d_{ij}$, where $s_{ij} \in [0; 1]$;
  - if $s_{ij} \in [0; 1]$, then $d_{ij} = 1 - s_{ij}$, where $d_{ij} \in [0; 1]$;
  - if $d_{ij} \in [0; \infty]$, then $s_{ij} = 1 - \frac{d_{ij} - \min_{i,j}(d_{ij})}{\max_{i,j}(d_{ij}) - \min_{i,j}(d_{ij})}$, where $s_{ij} \in [0; 1]$;
  - if $s_{ij} \in [0; \infty]$, then $d_{ij} = 1 - \frac{s_{ij} - \min_{i,j}(s_{ij})}{\max_{i,j}(s_{ij}) - \min_{i,j}(s_{ij})}$, where $d_{ij} \in [0; 1]$.

Semantic Similarity Measures

Definition

A semantic similarity measure quantifies semantic relatedness input terms c_i, c_j with the similarity score $s_{ij} = sim(c_i, c_j)$:

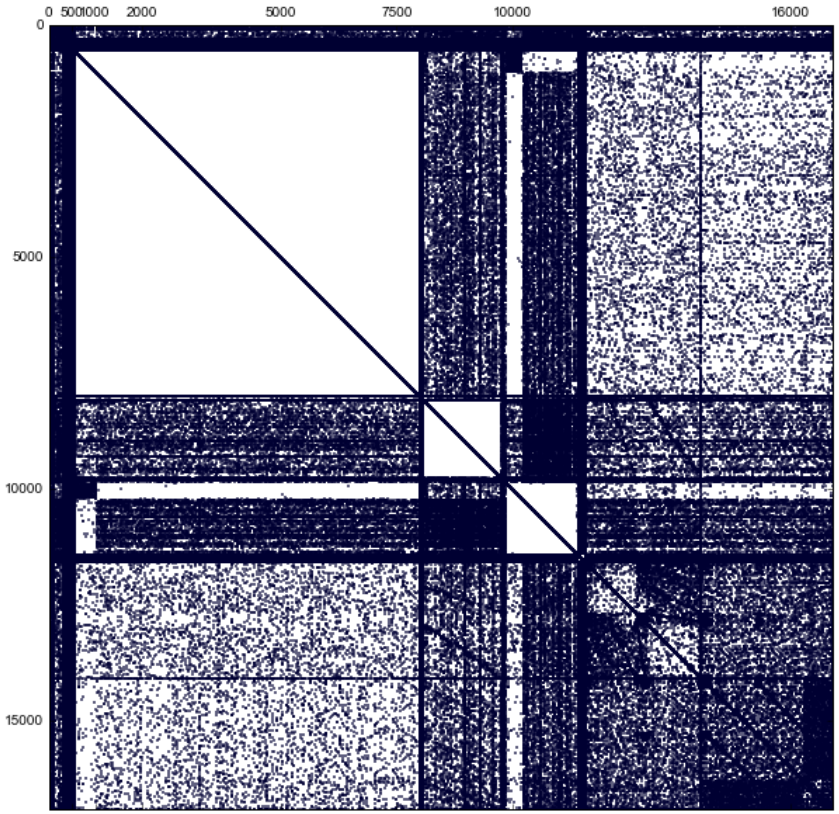
$$s_{ij} = \begin{cases} 1 & \text{if } \langle c_i, c_j \rangle \text{ is a pair of } syn, hyper, cohypo \\ 0 & \text{otherwise} \end{cases}$$

Properties

- Nonnegativity: $0 \leq s_{ij} \leq 1$;
- Reflexivity: $s_{ij} = 1 \Leftrightarrow c_i = c_j$;
- Symmetricity: $s_{ij} = s_{ji}$;

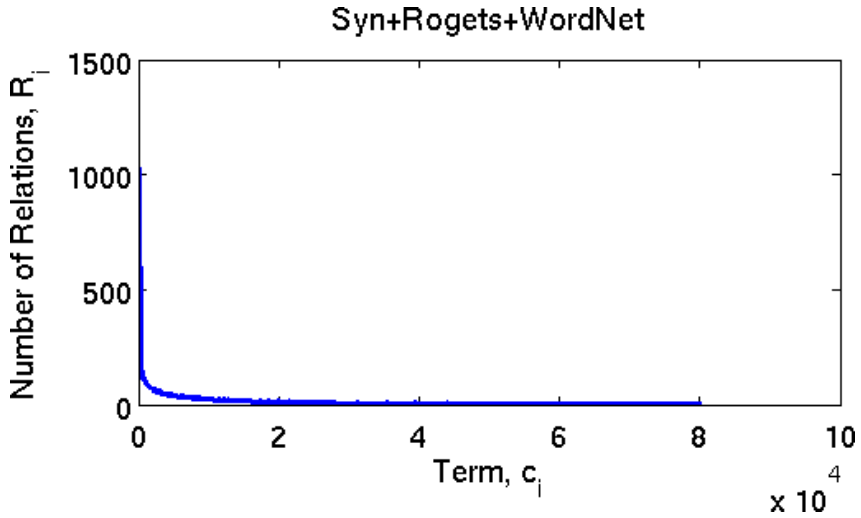
Word similarity matrix **S**

- **S** – word $*$ word similarity matrix;
- $s_{ij} \in \mathbf{S}$ – similarity of words w_i and w_j;
- $s_{ij} = sim(w_i, w_j), s_{ij} \in [0; 1]$.

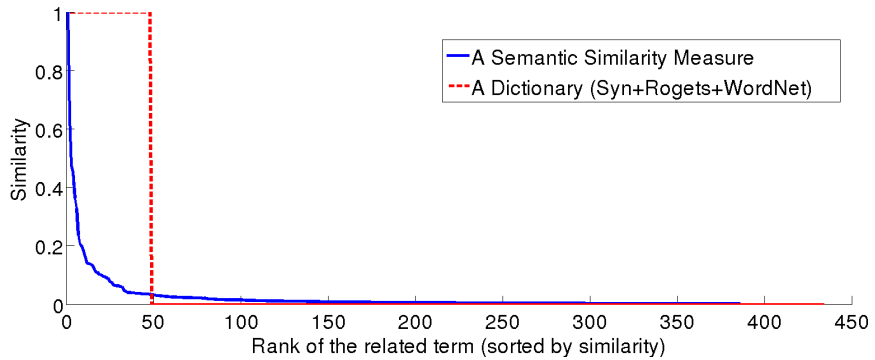


Semantic Similarity Measures

- Many dissimilar pairs, few similar pairs: $s_{ij} \sim exp(\lambda)$:



- Similarity distribution of the term "doctor":



Number of relations in semantic resources

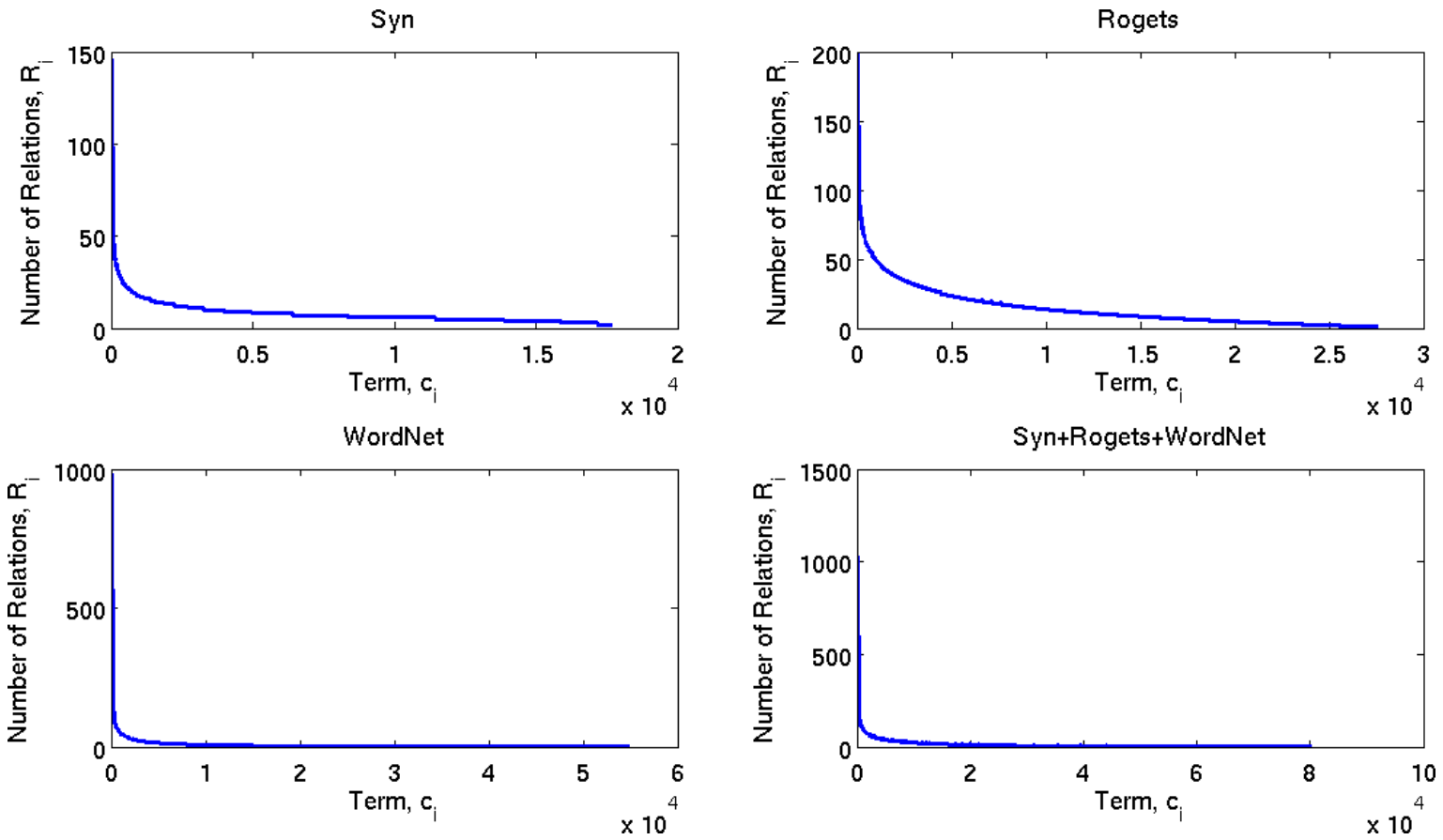


Figure: Number of relations (synonyms and hyponyms) per term in the dictionaries: a dictionary of synonyms, Roget's thesaurus, WordNet and a union of these three resources.

Evaluation of Semantic Similarity Measures

1. correlations with human judgments (**MC**, **RG**, **WordSim**);
2. semantic relation ranking (**BLESS**, **SN**);
3. semantic relation extraction;
4. using extracted relations in an application:
   - a short text classification system (**iCOP**);
   - a lexico-semantic search engine (**Serelex**).

Evaluation of Semantic Similarity Measures

1. **Correlations with human judgments**:
   - Criterion: Pearson correlation (ρ) ž Spearman correlation (r).
   - Datasets: MC, RG, WordSim.
2. **Semantic relation ranking**:
   - Criterion: Precision, Recall, F-measure.
   - Dataset: BLESS, SN.
3. **Semantic relation extraction:**
   - Criterion: Precision@k.
   - Data: annotation and/or dictionaries.
4. **Application-based evaluation:**
   - short text classification system (**iCOP**);
   - lexico-semantic search engine (**Serelex**).

Panchenko A., **Similarity Measures for Semantic Relation Extraction.** PhD thesis. Université catholique de Louvain. 197 pages, 2013, (Chapter 1).

Correlations with human judgments

word, c_i	word, c_j	human, **s**	sim, **s**	human (rank), **r**	sim (rank), $\hat{\mathbf{r}}$
tiger	cat	7.35	0.85	1	3
book	paper	7.46	0.95	2	2
computer	keyboard	7.62	0.81	3	1
...	...	...	...	...	...
possibility	girl	1.94	0.25	64	65
sugar	approach	0.88	0.05	65	23

Datasets:

- WordSim353 – 353 word pairs (Finkelstein, 2002)
- MC – 30 word pairs (Miller Charles, 1991)
- RG – 65 word pairs (Rubenstein Goodenough, 1965)

Pearson correlation: $\rho = \frac{cov(\mathbf{s},\hat{\mathbf{s}})}{\sigma(\mathbf{s})\sigma(\hat{\mathbf{s}})}$

Spierman correlation:: $r = \frac{cov(\mathbf{r},\hat{\mathbf{r}})}{\sigma(\mathbf{r})\sigma(\hat{\mathbf{r}})}$

Correlations with human judgments

Table 1.4: Miller-Charles (MC) dataset and scores obtained with a similarity measure.

Word, c_i	**Word**, c_j	**Human Score**, s_k	**Score**, $\hat{s}_k$	**Human Rank**, r_k	**Rank**, $\hat{r}_k$
automobile	car	3.92	0.884	1	1
journey	voyage	3.84	0.592	2	8
gem	jewel	3.84	0.581	3	3
boy	lad	3.76	0.325	4	2
coast	shore	3.70	0.440	5	7
asylum	madhouse	3.61	0.190	6	5
magician	wizard	3.50	0.556	7	4
midday	noon	3.42	0.692	8	10
furnace	stove	3.11	0.296	9	9
food	fruit	3.08	0.300	10	13
bird	cock	3.05	0.145	11	16
bird	crane	2.97	0.190	12	12
implement	tool	2.95	0.260	13	6
brother	monk	2.82	0.174	14	21
crane	implement	1.68	0.016	15	14
brother	lad	1.66	0.219	16	11
car	journey	1.16	0.124	17	25
monk	oracle	1.10	0.057	18	17
cemetery	woodland	0.95	0.056	19	24
food	rooster	0.89	0.027	20	26
coast	hill	0.87	0.186	21	28
forest	graveyard	0.84	0.069	22	23
shore	woodland	0.63	0.076	23	22
monk	slave	0.55	0.101	24	18
coast	forest	0.42	0.145	25	19
lad	wizard	0.42	0.083	26	20
cord	smile	0.13	0.020	27	29
glass	magician	0.11	0.078	28	27
noon	string	0.08	0.026	29	15
rooster	voyage	0.08	0.005	30	30

Semantic Relation Ranking

- Precision $P(k = 50) = \frac{1}{7} \approx 0.86$

word, c_i	**word,** c_j	**relation type**	s_{ij}
aficionado	enthusiast	syn	0.07197
aficionado	fan	syn	0.05195
aficionado	admirer	syn	0.01964
aficionado	addict	syn	0.01326
aficionado	devotee	syn	0.01163
aficionado	foundling	random	0.00777
aficionado	fanatic	syn	0.00414
aficionado	adherent	syn	0.00353
aficionado	capital	random	0.00232
aficionado	statute	random	0.00029
aficionado	blot	random	0.00025
aficionado	meddler	random	0.00005
aficionado	enlargement	random	0.00003
aficionado	bawdyhouse	random	0.00000