Universität Stuttgart

Multi-Word Measures: Modeling Semantic Change in Compound Nouns

Chris Jenkins
Filip Miletić
Sabine Schulte im Walde

Unrelated	"He never returned to his architectural **gold mine** of the prewar period."	"Next, there was a copra plantation up North, then a bogus **gold mine** in Colombia, a charter business,..."
Related	"One morning they saw a workman standing on the **brick wall**, who looked about him as if quite at leisure..."	"Most of the blast from a car bomb outside the French embassy was absorbed by a thick **brick wall**."

Summary

Contributions

- Dataset: (in-context) noun-compound diachronic relatedness ratings
- Expanding lexical-change detection task to complex lexical items
- Comparing vector representations, additional items included in (k-means) clustering
- Multiple evaluation perspectives: goodness of clustering, and degree of semantic change via internal and external measures

Takeaways

- Encouraging results from simpler representation
- Additional items in clustering of limited utility
- $|\Delta(L)|$, and PRT are best default measures
- Mixed result between languages using cluster structure (JSD)

System-Internal Change Measures

Prototype (PRT) — cosine-dist(avg( ) , avg( ))

Average Pairwise Distance (APD) — Σ (cosine_dist(,)) / n

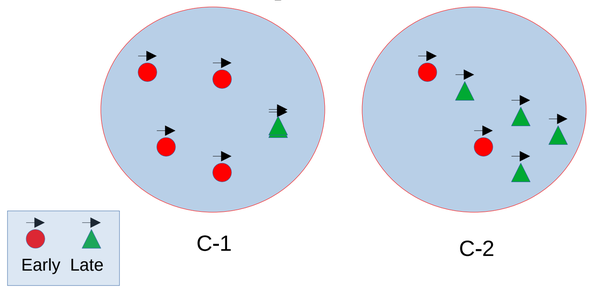
Jensen-Shannon Divergence (JSD)

$$D(x,y) = \sum x \, log\left(\frac{x}{y}\right)$$

$$m = \frac{(p+q)}{2}$$

$$JSD(p,q) = \frac{D(p,m) + D(q,m)}{2}$$

	C-1	C-2
Early	0.66	0.33
Late	0.20	0.80



System-External Change Measures

From pairwise **relatedness annotation**:
(1–4); 1: unrelated; 4: identical meanings

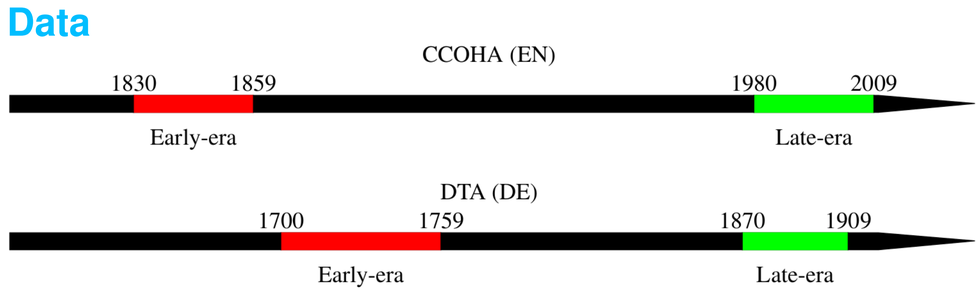
- $|\Delta(L)|$ Range 0.0–1.0; Difference in average relatedness ratings from late to early eras
- $\mu(C)$ Average rating for *compare* pairs: one from early, one from late era
- JSD_{anno} JSD of early and late distributions of clustered annotation graph

Additional Sources:

- **Compositionality Ratings** low: mod/head unrelated to compound; high: mod/head highly related to compound

Target	μ(E)	μ(L)	μ(C)	ΔL	JSD
field work	2.06	2.92	2.42	0.86	0.00
ins. company	2.77	3.52	3.21	0.76	0.13
nest egg	2.26	2.98	2.53	0.72	0.22
silver spoon	2.17	2.89	2.30	0.72	0.27
...	...	...	...	...	
fairy tale	2.61	2.36	2.73	-0.25	0.15
mother tongue	3.39	2.92	3.33	-0.47	0.00
love song	3.18	2.61	2.78	-0.57	0.24
...	...	...	...	...	
gold mine	2.88	1.94	2.49	-0.93	0.33
elbow room	2.71	1.77	2.84	-0.94	0.24
ground floor	3.57	2.59	3.33	-0.99	0.18
calendar month	3.81	2.62	2.73	-1.19	0.00
Ruhestand (*retirement*)	2.43	3.58	2.27	1.16	0.28
Rechtsstreit (*legal dispute*)	2.63	3.74	3.51	1.10	0.11
Uhrwerk (*clockwork*)	2.48	3.44	2.21	0.97	0.16
Triebwerk (*engine*)	2.22	3.14	1.88	0.92	0.41
...	...	...	...	...	
Sonnenstrahl (*sunbeam*)	2.88	2.9	3.18	0.02	0.07
Heerführer (*general (military)*)	3.37	3.36	3.57	-0.01	0.00
Sonnenlicht (*sunlight*)	3.01	2.96	3.04	-0.05	0.03
...	...	...	...	...	
Windspiel (*wind chimes*)	3.08	2.43	2.81	-0.65	0.17
Sonnenblume (*sunflower*)	3.34	2.70	3.23	-0.65	0.11
Salzwasser (*salt water*)	3.44	2.67	2.86	-0.76	0.21
Feldzug (*campaign*)	3.53	2.46	3.42	-1.07	0.13

Data

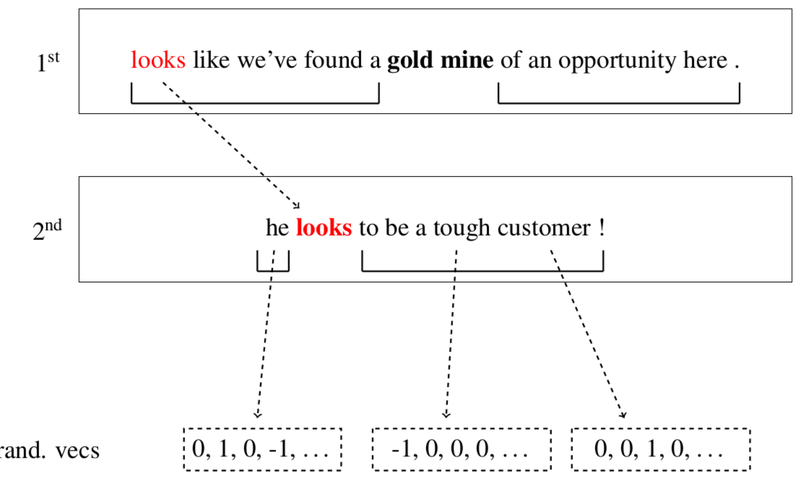


Corpus	Early Era	Tokens	Late Era	Tokens
CCOHA	1830–1859	51M	1980–2009	92M
DTA	1700–1759	33M	1870–1909	36M

Corpus	Compounds	Ex. Pairs	Ratings (skipped)	Avg. ρ Agr.
CCOHA	19	393	2702 (111)	0.33
DTA	43	899	5144 (308)	0.28

Representations

- **BERT** embeddings (domain-adapted on historical corpora), avg. of first four layers (avg. over multiple tokens when applicable)
- **2nd**-order random indexing vectors (see diagram below). Derived *exclusively* from historical corpora
- **Both** of the above representations, concatenated



Items Clustered

gold mine + { ∅ — *compounds*; gold × 4 — *modifier-constituent*; mine × 4 — *head-constituent*; **gold** chain × 1, **gold** dust × 1, ... — *mod-**c**onstituent **f**amily **s**et*; iron **mine** × 1, silver **mine** × 1, ... — *head-**c**onstituent **f**amily **s**et*

Results

Pairwise Accuracy

	en			de		
Setting	Min.	Max.	Avg.	Min.	Max.	Avg.
Vecs_{BERT}	0.40	0.44	0.42	0.28	0.31	0.29
Vecs_{2nd}	0.44	0.50	**0.47**	0.45	0.55	**0.51**
Vecs_{both}	0.40	0.47	0.43	0.27	0.30	0.28
$\text{targets}_{compounds}$	0.43	0.50	**0.46**	0.27	0.54	0.37
$\text{targets}_{head-const}$	0.42	0.46	0.44	0.28	0.47	0.35
$\text{targets}_{mod-const}$	0.40	0.49	0.43	0.30	0.55	**0.39**
$\text{targets}_{head-cfs}$	0.44	0.47	0.45	0.27	0.55	0.37
$\text{targets}_{mod-cfs}$	0.40	0.44	0.42	0.29	0.45	0.34

V-Measure

	en			de		
Setting	Min.	Max.	Avg.	Min.	Max.	Avg.
Vecs_{BERT}	0.26	0.39	**0.34**	0.21	0.38	0.33
Vecs_{2nd}	0.22	0.39	0.28	0.14	0.21	0.17
Vecs_{both}	0.26	0.39	0.33	0.23	0.39	**0.35**
$\text{targets}_{compounds}$	0.22	0.32	0.28	0.17	0.39	0.31
$\text{targets}_{head-const}$	0.27	0.36	0.33	0.21	0.38	**0.32**
$\text{targets}_{mod-const}$	0.22	0.37	0.32	0.17	0.37	0.29
$\text{targets}_{head-cfs}$	0.39	0.39	**0.39**	0.16	0.38	0.30
$\text{targets}_{mod-cfs}$	0.26	0.29	0.27	0.14	0.23	0.19

Semantic Change Evaluation

	$\mu(C)$		$\lvert\Delta(L)\rvert$		Compos.		JSD_{anno}	
	acc	vm	acc	vm	acc	vm	acc	vm
English								
PRT	-0.54^{*}	-0.70^{*}	0.47^{*}	0.08		-0.44	0.17	0.55
APD	-0.64^{**}	-0.70^{*}	0.32	0.23		-0.55	0.45	0.65
$JSD_{e,l}$	-0.50^{*}	-0.63	0.51^{*}	-0.05		-0.14	0.22	0.41
German								
PRT	0.08	0.07	0.37^{*}	0.22	0.35^{*}		0.03	0.05
APD	0.11	0.19	0.26	0.19	0.32^{*}		-0.05	-0.01
$JSD_{e,l}$	-0.12	-0.01	0.18	0.14	0.20		-0.11	0.15

Spearman's ρ correlations between internal (rows) and external (columns) semantic change measures. $*: p < 0.05$, $**: p < 0.01$ Best configs per language in terms of accuracy (acc) and V-measure (vm):

Accuracy:

- English: best k:4, 2nd-order, compounds.
- German: best k:4, 2nd-order, head-cfs;

V-measure:

- English: best k:8, 2nd-order, head-cfs.
- German: best k:30, both, compounds

"The use of the word 'meaning' is subject to the general rule that each word when used in a new context is a new word." (Firth, 1957, p.190)