A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains

Dominik Schlechtweg[1], Anna Hätty[1,2], Marco del Tredici[3], Sabine Schulte im Walde[1]
[1]Institute for Natural Language Processing, University of Stuttgart
[2]Robert Bosch GmbH, Corporate Research
[3]Institute for Logic, Language and Computation, University of Amsterdam
{schlecdk,schulte}@ims.uni-stuttgart.de,
anna.haetty@de.bosch.com, m.deltredici@uva.nl

Diachronic LSC Detection

- Semantic vector spaces
  - Each word as vector1 and vector2, measured by cosine-distance or other metrics
- Topic distributions
  - Infer different word senses(topics)
- Sense clusters
  - Similar to above one.

Synchronic LSC Detection

- focus on how the meanings of words vary across domains or communities of speakers.

Evaluation

- Empirically observed data.
- Synthetic data or related tasks.

Data

- DTA18
- DTA19
- SDEWAC
- COOK

	Times		Domains	
	DTA18	DTA19	SDEWAC	COOK
L_{ALL}	26M	40M	109M	1M
L/P	10M	16M	47M	0.6M

Table 1: Corpora and their approximate sizes.

- DURel (Diachronic Usage Relatedness)
  - 22 target words
  - 1750-1799 / 1850-1899
- SURel (Synchronic Usage Relatedness)
  - 22 target words
  - Cooking recipes / general language

Meaning Representations

- Semantic Vector Spaces
  - Count-based Vector Spaces
  - Predictive Vector Spaces
  - Alignment
- Topic Distributions
  - Sense Change

Count-based Vector Spaces

- Positive Pointwise Mutual Information

$$M_{i,j}^{\text{PPMI}} = \max \left\{ \log \left(\frac{\#(w_i, c_j) \sum_c \#(c)^\alpha}{\#(w_i)\#(c_j)^\alpha} \right) - \log(k), 0 \right\}$$

k > 1 is a prior on the probability of observing an actual occurrence of (w_i, c_j).

0 < α < 1 is a smoothing parameter reducing PPMI's bias towards rare words.

Count-based Vector Spaces

- Singular Value Decomposition

$$M_d = U_d \cdot \Sigma_d \cdot V_d^\top$$

$$M^{\mathrm{SVD}} = U_d \Sigma_d^p$$

p is an eigenvalue weighting parameter
The *i_th* row of M^{SVD} corresponds to w_i's d-dimensional representation.

Count-based Vector Spaces

- Random Indexing

$$M^{\mathrm{RI}} = MR^{|\mathcal{V}|\times d}$$

- points in a vector space can be mapped into a randomly selected subspace under approximate preservation of the distances between points, if the subspace has a sufficiently high dimensionality.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2015. Temporal random indexing: A system for analysing word meaning over time. Italian Journal ofComputational Linguistics, 1:55–68.

Predictive Vectors Spaces

- Skip-Gram with Negative Sampling (SGNS)

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c)\in D'} \log \sigma(-v_c \cdot v_w),$$

  - D is the set of all observed word-context pairs
  - D’ is the set of randomly generated negative samples and is obtained by drawing k contexts from the empirical unigram distribution

Alignment

- Column Intersection

$$A^{\mathrm{CI}}_{*j} = A_{*j} \quad \text{for all } c_j \in V_a \cap V_b,$$
$$B^{\mathrm{CI}}_{*j} = B_{*j} \quad \text{for all } c_j \in V_a \cap V_b,$$

X_{*j} denotes the jth column of X.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th Annual Meeting ofthe Association for Computational Linguistics, pages 1489–1501, Berlin, Germany.

Alignment

- Shared Random Vector (SVR)

$$A^{\mathrm{SVR}} = AR,$$
$$B^{\mathrm{SVR}} = BR.$$

Count matrices A and B are multiplied both by the same random matrix R representing them in the same low-dimensional random space.

Alignment

- Orthogonal Procrustes (OP)

$$W^* = \arg\min_W \sum_i \sum_j D_{i,j} \|B_{i*}W - A_{j*}\|^2$$

$$A^{\mathrm{OP}} = A,$$

$$B^{\mathrm{OP}} = BW^*$$

D is a binary matrix represents the dictionary, $D_{i,j} = 1$ if w_i in the vocabulary at time b.

https://link.springer.com/chapter/10.1007/978-3-030-11760-3_2

Alignment

- Vector Initialization (VI)

1. Use standard SGNS to learn A^{VI}
2. Initialize the SGNS model for learning B^{VI} on A^{VI}

If a word is used in similar contexts in a and b, its vector will be updated only slightly, while more different contexts lead to a stronger update.

Alignment

- Word Injection (WI)

B->B' by 'walk' -> '_walk'

Use mixed corpus A+B' to obtain a single matrix

Ferrari, A., Donati, B., & Gnesi, S. Detecting domain-specific ambiguities: an NLP approach based on Wikipedia crawling and word embeddings. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)* (pp. 393-399). IEEE.

LSC Detection Measures

- Similarity Measures
  - Similarity Distance (**CD**)

$$cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\sqrt{\vec{x}\cdot\vec{x}}\sqrt{\vec{y}\cdot\vec{y}}}$$

$$CD(\vec{x},\vec{y}) = 1 - cos(\vec{x},\vec{y})$$

  - Local Neighborhood Distance (**LND**)

$$s(j) = \cos(\vec{x},\vec{z}_j) \quad \forall \vec{z}_j \in N_k(\vec{x}) \cup N_k(\vec{y})$$

$$LND(\vec{x},\vec{y}) = CD(\vec{s_x},\vec{s_y})$$

  - Jensen-Shannon Distance (**JSD**)

$$M = (\phi_x + \phi_y)/2$$

$$JSD(\phi_x||\phi_y) = \sqrt{\frac{D_{KL}(\phi_x||M) + D_{KL}(\phi_y||M)}{2}}$$

LSC Detection Measures

- Dispersion Measures
  - Frequency Difference (**FD**)

$$F(w, C) = \log \frac{|w \in C|}{|C|} \qquad FD(x, X, y, Y) = |F(x, X) - F(y, Y)|$$

  - Type Difference (**TD**)

$$T(\vec{w}, C) = \log \frac{\sum_{i=1} 1 \quad \text{if } \vec{w}_i \neq 0}{|C_T|} \quad TD(\vec{x}, X, \vec{y}, Y) = |T(\vec{x}, X) - T(\vec{y}, Y)|$$

  - Entropy Difference (**HD**)

$$VH(\vec{w}) = -\sum_{i=1} \frac{\vec{w}_i}{\sum_{j=1} \vec{w}_j} \log \frac{\vec{w}_i}{\sum_{j=1} \vec{w}_j} \quad HD(\vec{x}, \vec{y}) = |VH(\vec{x}) - VH(\vec{y})|$$

Result

Dataset	Representation	best	mean
DURel	raw count	0.639	0.395
	PPMI	0.670	0.489
	SVD	0.728	0.498
	RI	0.601	0.374
	SGNS	**0.866**	**0.502**
	SCAN	0.327	0.156
SURel	raw count	0.599	0.120
	PPMI	0.791	0.500
	SVD	0.639	0.300
	RI	0.622	0.299
	SGNS	**0.851**	**0.520**
	SCAN	0.082	-0.244

Table 3: Best and mean ρ scores across similarity measures (CD, LND, JSD) on semantic representations.

Result

Dataset	Preproc	Win	Space	Parameters	Align	Measure	Spearman m (h, l)
	L_{ALL}	10	SGNS	k=1,t=None	OP	CD	**0.866** (0.914, 0.816)
	L_{ALL}	10	SGNS	k=5,t=None	OP	CD	0.857 (0.891, 0.830)
DURel	L_{ALL}	5	SGNS	k=5,t=0.001	OP	CD	0.835 (0.872, 0.814)
	L_{ALL}	10	SGNS	k=5,t=0.001	OP	CD	0.826 (0.863, 0.768)
	L/P	2	SGNS	k=5,t=None	OP	CD	0.825 (0.826, 0.818)
	L/P	2	SGNS	k=1,t=0.001	OP	CD	**0.851** (0.851, 0.851)
	L/P	2	SGNS	k=5,t=None	OP	CD	0.850 (0.850, 0.850)
SURel	L/P	2	SGNS	k=5,t=0.001	OP	CD	0.834 (0.838, 0.828)
	L/P	2	SGNS	k=5,t=0.001	OP_{-}	CD	0.831 (0.836, 0.817)
	L/P	2	SGNS	k=5,t=0.001	OP	CD	0.829 (0.832, 0.823)

Table 2: Best results of ρ scores (Win=Window Size, Preproc=Preprocessing, Align=Alignment, k=negative sampling, t=subsampling, Spearman m(h,l): mean, highest and lowest results).

Result

Dataset	OP	OP_{-}	OP_{+}	WI	None
DURel	0.618	0.557	**0.621**	0.468	0.254
SURel	**0.590**	0.514	0.401	0.492	0.285

Table 4: Mean ρ scores for CD across the alignments. Applies only to RI, SVD and SGNS.

Thanks

- Can we use pre-trained language models and how ?