Modeling lexical semantic shifts during ad-hoc coordination

Alexandre Kabbach[1,2] Aurélie Herbelot[2]

18.05.2020 – GeCKo 2020

[1]University of Geneva

[2]CIMeC – University of Trento

Problem

Speakers form conceptual representations for words based on different *background experiences* (Connell and Lynott, 2014).

Conceptual variability and communication

Speakers form conceptual representations for words based on different *background experiences* (Connell and Lynott, 2014).

How can speakers nonetheless communicate with one another if the words they utter do not refer to the exact same concepts?

Coordination: a possible solution?

Speakers *coordinate* with one-another during each communication instance in order to settle for specific word meanings (Clark, 1992, 1996).

In doing so, they *contextualize* their *generic* conceptual representations during communication.

Question

How can we integrate coordination to standard Distributional Semantic Models (DSMs; Turney and Pantel, 2010; Clark, 2012; Erk, 2012; Lenci, 2018)?

Problems:

1. DSMs do not distinguish background linguistic stimuli from active coordination in their acquisition process
2. DSMs consider conceptual representations to remain invariant during communication

Proposal

Model

We distinguish *background experience* from *ad-hoc coordination* in a standard count-based PPMI-weighted DSM:

Model

We distinguish *background experience* from *ad-hoc coordination* in a standard count-based PPMI-weighted DSM:

- **background experience** = corpus data fed to the DSM

Model

We distinguish *background experience* from *ad-hoc coordination* in a standard count-based PPMI-weighted DSM:

- **background experience** = corpus data fed to the DSM
- **ad-hoc coordination** = singular vector sampling in the SVD

Model

We distinguish *background experience* from *ad-hoc coordination* in a standard count-based PPMI-weighted DSM:

- **background experience** = corpus data fed to the DSM
- **ad-hoc coordination** = singular vector sampling in the SVD

We replace the variance-preservation bias in the SVD of the DSM by an explicit coordination bias, sampling the set of d singular vectors which maximize the correlation with a particular similarity dataset (MEN and SimLex).

Assumptions

Assumptions

1. a single DSM can capture different kinds of semantic relations from the same corpus, so that a collection of possible meaning spaces could coexist within the same set of data

Assumptions

1. a single DSM can capture different kinds of semantic relations from the same corpus, so that a collection of possible meaning spaces could coexist within the same set of data
2. aligning similarity judgments across sets of word pairs provides a nice approximation of ad-hoc coordination between two speakers originally disagreeing and ultimately converging to a form of agreement with respect to some lexical decision

Results

Results

1. replacing the variance preservation bias with an explicit sampling bias actually *reduces the variability* across models generated from different corpora

Results

1. replacing the variance preservation bias with an explicit sampling bias actually *reduces the variability* across models generated from different corpora
2. DSMs generated from different corpora can be aligned in different ways. Alignment does not necessarily equate conceptual *agreement* but in some cases, mere *compatibility*, so that coordinating one's conceptual spaces might simply be the cooperative act of *avoiding conflict*, rather than being in full agreement

Model

PPMI-weighted DSM

$$PMI(w,c) = log\frac{P(w,c)}{P(w) \cdot P(c)}$$

$$PPMI = max(PMI(w,c), 0)$$

$$W = U \cdot \Sigma \cdot V^{\top}$$

$$W_d = U_d \cdot \Sigma_d^{\alpha} \quad \alpha \in [0,1]$$

Singular vector sampling

$$W_d = U_d \cdot \Sigma_d^{\alpha} \quad \alpha \in [0, 1]$$

Singular vector sampling

$$W_d = U_d \cdot \Sigma_d^{\alpha} \quad \alpha \in [0, 1]$$

Replace the variance-preservation bias by the following add-reduce algorithm:

Singular vector sampling

$$W_d = U_d \cdot \Sigma_d^{\alpha} \quad \alpha \in [0, 1]$$

Replace the variance-preservation bias by the following add-reduce algorithm:

- **add**: iterate over all singular vectors and selects only those that increase performance on a given lexical similarity dataset

Singular vector sampling

$$W_d = U_d \cdot \Sigma_d^{\alpha} \quad \alpha \in [0, 1]$$

Replace the variance-preservation bias by the following add-reduce algorithm:

- **add**: iterate over all singular vectors and selects only those that increase performance on a given lexical similarity dataset
- **reduce**: iterate over the set of added singular vectors and removes all those that do not negatively alter performance on the given lexical similarity dataset

Conceptual similarity

We model *structural similarity* between two DSMs as the minimized Root Mean Square Error (RMSE) between them.

Conceptual similarity

We model *structural similarity* between two DSMs as the minimized Root Mean Square Error (RMSE) between them.

$$RMSE(A, B) = \sqrt{\frac{1}{|A|} \sum_{i=1}^{|A|} ||a_i - b_i||^2}$$

Conceptual similarity

We model *structural similarity* between two DSMs as the minimized Root Mean Square Error (RMSE) between them.

$$RMSE(A, B) = \sqrt{\frac{1}{|A|} \sum_{i=1}^{|A|} ||a_i - b_i||^2}$$

Models are aligned using *absolute orientation with scaling* (Dev et al., 2018) which minimizes the RMSE while applying cosine similarity-preserving linear transformation (rotation + scaling).

Experimental setup: corpora

Corpus	Word Count	Details
OANC	17M	Open American National Corpus
WIKI07	19M	.7% of the English Wikipedia
ACL	58M	ACL anthology reference corpus
WIKI2	53M	2% of the English Wikipedia
BNC	113M	British National Corpus
WIKI4	106M	4% of the English Wikipedia
WIKI	2 600M	Full English Wikipedia of January 20 2019

Table 1: Corpora used to generate DSMs

Experimental setup: lexical similarity

Experimental setup: lexical similarity

1. MEN (Bruni et al., 2014) *relatedness* dataset containing 3 000 word pairs. Expresses topical association (i.e. *cat* and *meow* are deemed related)

Experimental setup: lexical similarity

1. MEN (Bruni et al., 2014) *relatedness* dataset containing 3 000 word pairs. Expresses topical association (i.e. *cat* and *meow* are deemed related)
2. SimLex-999 (Hill et al., 2015) *similarity* dataset containing 999 word pairs. Expresses categorical similarity (i.e. *cat* and *dog* might be considered similar in virtue of being members of the same category)

Experimental setup: lexical similarity

1. MEN (Bruni et al., 2014) *relatedness* dataset containing 3 000 word pairs. Expresses topical association (i.e. *cat* and *meow* are deemed related)
2. SimLex-999 (Hill et al., 2015) *similarity* dataset containing 999 word pairs. Expresses categorical similarity (i.e. *cat* and *dog* might be considered similar in virtue of being members of the same category)

Those two datasets encode possibly incompatible semantic constraints and it is theoretically impossible to perfectly fit both the meaning spaces they encode with a single DSM (e.g. "chicken-rice" has a similarity score of 0.68 in MEN and 0.14 in SimLex).

Results

No variance-preservation bias means better DSMs

	WIKI07	OANC	WIKI2	ACL	WIKI4	BNC	WIKI
SVD-TOP ($\alpha = 1$)	0.61	0.60	0.66	0.26	0.66	0.70	0.67
SVD-TOP ($\alpha = 0$)	0.65	0.66	0.70	0.37	0.72	0.75	0.74
SVD-SEQ	0.65 ± 0.02	0.66 ± 0.01	0.70 ± 0.02	**0.55** ± 0.02	0.71 ± 0.01	0.76 ± 0.01	0.76 ± 0.00

Table 2: Spearman correlation on MEN for DSMs generated from different corpora. SVD-TOP are PPMI-weighted count-based models reduced by selecting the top 300 singular vectors, with ($\alpha = 1$) or without ($\alpha = 0$) singular values. SVD-SEQ results are generated via our sampling algorithm and averaged across test sets applying 5-fold validation

No variance-preservation bias means better DSMs

	WIKI07	OANC	WIKI2	ACL	WIKI4	BNC	WIKI
SVD-TOP ($\alpha = 1$)	0.27	0.19	0.30	0.10	0.31	0.31	0.31
SVD-TOP ($\alpha = 0$)	0.31	0.23	0.34	0.15	0.36	0.37	0.37
SVD-SEQ	0.27 $\pm$ 0.08	0.22 $\pm$ 0.06	0.32 $\pm$ 0.03	**0.24** $\pm$ 0.04	0.36 $\pm$ 0.05	0.40 $\pm$ 0.07	**0.44** $\pm$ 0.05

Table 3: Spearman correlation on SimLex for DSMs generated from different corpora. SVD-TOP are PPMI-weighted count-based models reduced by selecting the top 300 singular vectors, with ($\alpha = 1$) or without ($\alpha = 0$) singular values. SVD-SEQ results are generated via our sampling algorithm and averaged across test sets applying 5-fold validation

No variance-preservation bias means more compact DSMs

	WIKI07	OANC	WIKI2	ACL	WIKI4	BNC	WIKI
SVD-TOP	300	300	300	300	300	300	300
SVD-SEQ-MEN	124 ± 10	175 ± 8	130 ± 7	308 ± 21	175 ± 11	128 ± 8	198 ± 16
SVD-SEQ-SIMLEX	55 ± 9	216 ± 21	121 ± 8	205 ± 29	136 ± 10	133 ± 11	185 ± 6

Table 4: Comparing dimensionality (number of selected singular vectors) between TOP and SEQ models. Dimensionality for SEQ models is averaged across 5-fold test sets results

Different dimensions encode different semantic phenomena

	MEN			SimLex		
	median	mean	90%	median	mean	90%
WIKI07	103 ± 16	845 ± 216	2653 ± 1363	595 ± 257	2012 ± 366	6454 ± 787
OANC	135 ± 31	687 ± 163	1803 ± 930	905 ± 403	2274 ± 487	6921 ± 1146
WIKI2	117 ± 15	687 ± 119	1285 ± 1071	390 ± 117	1515 ± 234	5471 ± 861
ACL	601 ± 53	1205 ± 107	2981 ± 445	910 ± 80	1925 ± 122	5842 ± 701
WIKI4	119 ± 13	426 ± 113	626 ± 143	398 ± 76	1290 ± 185	4321 ± 93
BNC	110 ± 22	436 ± 179	843 ± 448	394 ± 59	1280 ± 104	3810 ± 525
WIKI	185 ± 41	513 ± 135	1023 ± 318	657 ± 108	1259 ± 160	3160 ± 69

Table 5: Average mean, median and 90-th percentile of sampled dimensions indexes on MEN and SimLex for 10 shuffled runs

Coordination is an interactive process

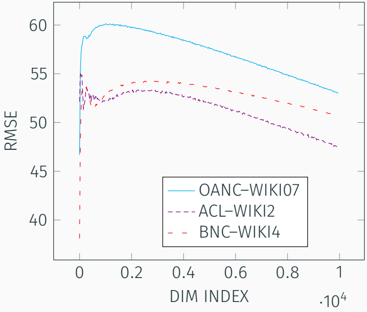

Figure 1: Evolution of RMSE for aligned bins of 30 consecutive singular vectors sampled across [0, 10 000] for aligned corpora of different domains but similar size.

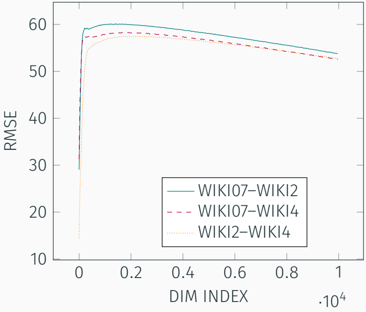

Figure 2: Evolution of RMSE for aligned bins of 30 consecutive singular vectors sampled across [0, 10 000] for aligned corpora of similar domains but different size.

Agreement *versus* compatibility

Two given models may be aligned if they both have *similar* components, but also if they have *dissimilar* components, provided that those components do not *conflict*.

Agreement *versus* compatibility

Two given models may be aligned if they both have *similar* components, but also if they have *dissimilar* components, provided that those components do not *conflict*.

Notions of *agreement*, *compatibility* and *conflict* can be defined via the absolute Pearson correlation r. Example:

Agreement *versus* compatibility

Two given models may be aligned if they both have *similar* components, but also if they have *dissimilar* components, provided that those components do not *conflict*.

Notions of *agreement*, *compatibility* and *conflict* can be defined via the absolute Pearson correlation r. Example:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} .9 & 0 & 0 & 0 \\ 0 & .9 & 0 & 0 \\ 0 & 0 & .9 & 0 \\ 0 & 0 & 0 & .9 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

- $RMSE(A, B) \sim RMSE(B, C) \sim RMSE(A, C) \approx 0$; but
- $r(A, B) = 1$ while $r(A, C) = 0.3$

Beyond similarity: conceptual compatibility

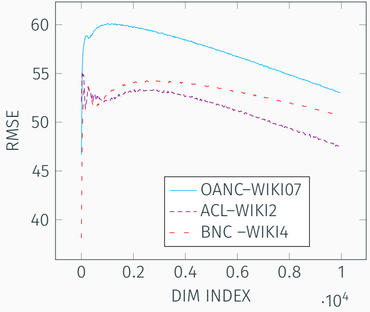

Figure 3: Evolution of RMSE for aligned bins of 30 consecutive singular vectors sampled across [0, 10 000] for aligned corpora of different domains but similar size.

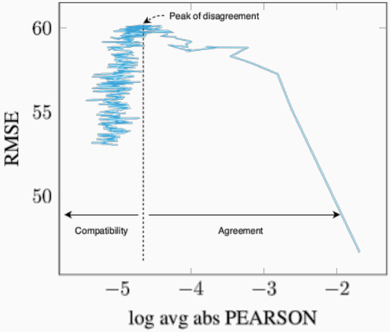

Figure 4: Evolution of RMSE with log of average absolute PEARSON correlation for aligned bins of 30 consecutive singular vectors sampled across [0, 10 000] on OANC and WIKI07.

Summary

Summary

1. replacing the variance preservation bias with an explicit sampling bias actually *reduces the variability* across models generated from different corpora

Summary

1. replacing the variance preservation bias with an explicit sampling bias actually *reduces the variability* across models generated from different corpora
2. DSMs generated from different corpora can be aligned in different ways. Alignment does not necessarily equate conceptual *agreement* but in some cases, mere *compatibility*, so that coordinating one's conceptual spaces might simply be the cooperative act of *avoiding conflict*, rather than being in full agreement

Summary

1. replacing the variance preservation bias with an explicit sampling bias actually *reduces the variability* across models generated from different corpora
2. DSMs generated from different corpora can be aligned in different ways. Alignment does not necessarily equate conceptual *agreement* but in some cases, mere *compatibility*, so that coordinating one's conceptual spaces might simply be the cooperative act of *avoiding conflict*, rather than being in full agreement
3. the number of *compatible* subspaces across the SVD largely extend the number of *agreeing* ones, so that speakers can never be expected to *agree* more than to some extent

Questions?

Cognitive plausibility 1/3

- DSMs stand in the long tradition of learning theories which argue that humans are excellent in capturing statistical regularities in their environments (Anderson and Schooler, 1991)
- PPMI-based weighting captures informativity between words and contexts rather than raw co-occurrence counts, and this fact is also in line with learning theories that emphasize that *contingency*, not *contiguity*, drives learning of associations between stimuli (Rescorla and Wagner, 1972; Murdock, 1982)

Cognitive plausibility 2/3

- Dimensionality reduction in DSMs models the transition from episodic to semantic memory, formalized as the generalization of observed concrete instances of word-context co-occurrences to higher-order representations potentially capturing more fundamental and conceptual relations (Landauer and Dumais, 1997)
- Humans apply dimensionality reduction as a *data compression* mechanism in order to facilitate encoding, memory and overall processing (Edelman, 1999)

Cognitive plausibility 3/3

- Cognitive plausibility of transformational alignment-based similarity is more delicate, for we merely use it as an approximation to serve as a proxy for modeling coordination. Two speakers will never gain access to each other's conceptual space, and as such the minimization of the RMSE between two DSMs remains a conceptual tool which has no psychological reality