Distributional Semantic Models

Part 3: Evaluation of distributional similarity

Stefan Evert[1]
with Alessandro Lenci[2], Marco Baroni[3] and Gabriella Lapesa[4]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]University of Pisa, Italy
[3]University of Trento, Italy
[4]University of Stuttgart, Germany

http://wordspace.collocations.de/doku.php/course:start

Outline

Outline

Distributional similarity as semantic similarity

- DSMs interpret semantic similarity as a quantitative notion
  - if **a** is closer to **b** than to **c** in the distributional vector space, then *a* is more semantically similar to *b* than to *c*
- Different from categorical nature of most theoretical accounts
  - often expressed in terms of semantic classes and relations
- But it is not clear a priori what exactly makes two words or concepts "semantically similar" according to a DSM
  - may also depend on parameter settings

Types of semantic relations in DSMs

Nearest DSM neighbors have different types of semantic relations.

car (BNC, L2/R2 span)

- van co-hyponym
- vehicle hyperonym
- truck co-hyponym
- motorcycle co-hyponym
- driver related entity
- motor part
- lorry co-hyponym
- motorist related entity
- cavalier hyponym
- bike co-hyponym

car (BNC, L30/R30 span)

- drive function
- park typical action
- bonnet part
- windscreen part
- hatchback part
- headlight part
- jaguar hyponym
- garage location
- cavalier hyponym
- tyre part

http://clic.cimec.unitn.it/infomap-query/

Manual analysis of semantic relations

for 44 concrete English nouns (Baroni & Lenci 2008)

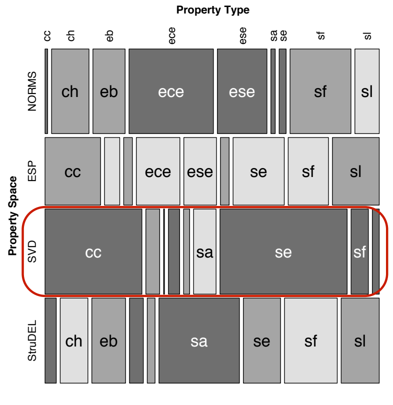


Figure 1: Distribution of property types across property spaces.

Taxonomic category:
- cc (co-)hyponym
- ch hypernym

Properties of entity:
- eb typical behaviour
- ece external component
- ese surface property

Situationally associated:
- sa action
- se other entity
- sf function
- sl location

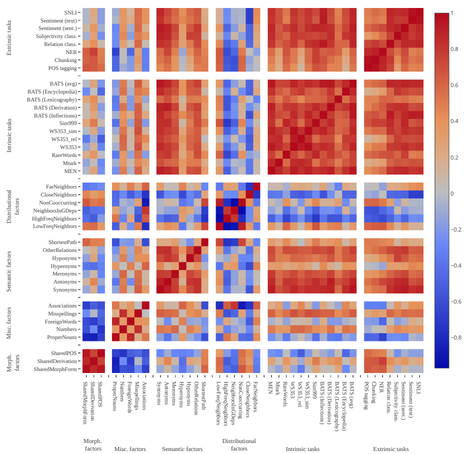
Distribution of semantic relations for top-10 L5/R5 DSM neighbours (SVD), pattern collocations (StruDEL) and human-generated properties (NORMS).

Scaling up: Linguistic Diagnostics

Linguistic Diagnostics (Rogers *et al.* 2018) automates classification of nearest neighbours based on various on-line dictionaries and semantic networks

- ☞ correlation analysis between various groups of diagnostics and evaluation tasks
- ☞ only for English so far

http://ldtoolkit.space/



Semantic similarity and relatedness

- Attributional similarity – two words sharing a large number of salient features (attributes)
  - synonymy (*car/automobile*)
  - hyperonymy (*car/vehicle*)
  - co-hyponymy (*car/van/truck*)
- Semantic relatedness (Budanitsky & Hirst 2006) – two words semantically associated without necessarily being similar
  - function (*car/drive*)
  - meronymy (*car/tyre*)
  - location (*car/road*)
  - attribute (*car/fast*)
- Relational similarity (Turney 2006) – similar relation between pairs of words (analogy)
  - *policeman* : *gun* :: *teacher* : *book*
  - *mason* : *stone* :: *carpenter* : *wood*
  - *traffic* : *street* :: *water* : *riverbed*

Outline

DSMs and semantic similarity

- DSMs are thought to represent paradigmatic similarity
  - words that tend to occur in the same contexts
- Words that share many contexts will correspond to concepts that share many attributes (attributional similarity), i.e. concepts that are taxonomically/ontologically similar
  - synonyms (*rhino/rhinoceros*)
  - antonyms and values on a scale (*good/bad*)
  - co-hyponyms (*rock/jazz*)
  - hyper- and hyponyms (*rock/basalt*)
- Taxonomic similarity is seen as the fundamental semantic relation organising the vocabulary of a language, allowing categorization, generalization and inheritance

Evaluation of attributional similarity

- Synonym identification
  - TOEFL test (Landauer & Dumais 1997)
- Modeling semantic similarity judgments
  - RG norms (Rubenstein & Goodenough 1965)
  - WordSim-353 (Finkelstein *et al.* 2002)
  - MEN (Bruni *et al.* 2014), SimLex-999 (Hill *et al.* 2015)
- Noun categorization
  - ESSLLI 2008 dataset
  - Almuhareb & Poesio (AP, Almuhareb 2006)
- Semantic priming
  - Hodgson dataset (Padó & Lapata 2007)
  - Semantic Priming Project (Hutchison *et al.* 2013)
- Analogies & semantic relations (similarity vs. relatedness)
  - Google (Mikolov *et al.* 2013b), BATS (Gladkova *et al.* 2016)
  - BLESS (Baroni & Lenci 2011), CogALex (Santus *et al.* 2016)

Give it a try . . .

- The `wordspace` package contains pre-compiled DSM vectors
  - based on a large Web corpus (9 billion words)
  - L4/R4 surface span, log-transformed G^2, SVD dim. red.
  - targets = lemma + POS code (e.g. `white_J`)
  - compatible with evaluation tasks included in package

```
library(wordspace)

M <- DSM_Vectors
nearest.neighbours(M, "walk_V")
    amble_V    stroll_V   traipse_V    potter_V     tramp_V
       19.4        21.8        21.8        22.6        22.9
  saunter_V    wander_V    trudge_V leisurely_R   saunter_N
       23.5        23.7        23.8        26.2        26.4
# you can also try white, apple and kindness
```

The TOEFL synonym task

- The TOEFL dataset (80 items)
  - Target: *levied*
    Candidates: *believed*, *correlated*, *imposed*, *requested*
  - Target *fashion*
    Candidates: *craze*, *fathom*, *manner*, *ration*
- DSMs and TOEFL
  1. take vectors of the target ($\mathbf{t}$) and of the candidates ($\mathbf{c}_1 \dots \mathbf{c}_n$)
  2. measure the distance between $\mathbf{t}$ and $\mathbf{c}_i$, with $1 \leq i \leq n$
  3. select $\mathbf{c}_i$ with the shortest distance in space from $\mathbf{t}$

```
# ask your course instructor for non-public data package
> library(wordspaceEval)
> head(TOEFL80)
```


Humans vs. machines on the TOEFL task

- Average foreign test taker: 64.5%
- Macquarie University staff (Rapp 2004):
  - Average of 5 non-natives: 86.75%
  - Average of 5 natives: 97.75%
- Distributional semantics
  - Classic LSA (Landauer & Dumais 1997): 64.4%
  - Padó and Lapata's (2007) dependency-based model: 73.0%
  - Distributional memory (Baroni & Lenci 2010): 76.9%
  - Rapp's (2004) SVD-based model, lemmatized BNC: 92.5%
  - Bullinaria & Levy (2012) carry out aggressive parameter optimization: 100.0%

And you?

```
> eval.multiple.choice(TOEFL80, M)
```


Semantic similarity judgments

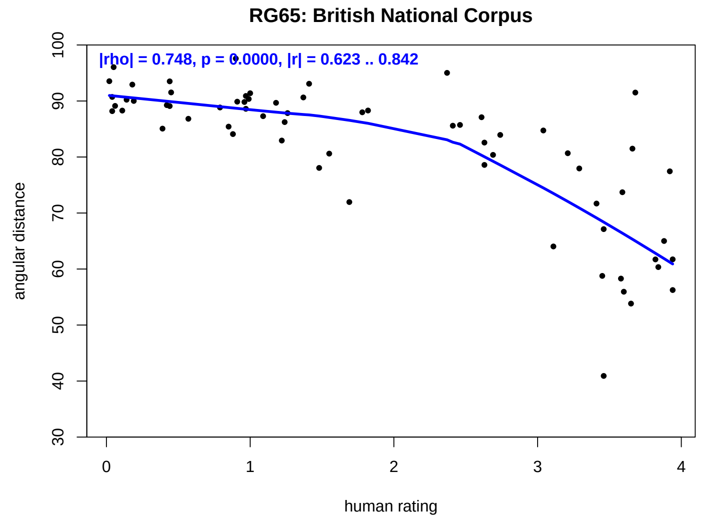
- Rubenstein & Goodenough (1965) collected similarity ratings for 65 noun pairs from 51 subjects on a 0–4 scale

w_1	w_2	avg. rating
car	*automobile*	3.9
food	*fruit*	2.7
cord	*smile*	0.0

- DSMs vs. Rubenstein & Goodenough
  1. for each test pair (w_1, w_2), take vectors $\mathbf{w}_1$ and $\mathbf{w}_2$
  2. measure the distance (e.g. cosine) between $\mathbf{w}_1$ and $\mathbf{w}_2$
  3. measure (Pearson) correlation between vector distances and R&G average judgments (Padó & Lapata 2007)

```
> RG65[seq(0,65,5), ]
> head(WordSim353) # extension of Rubenstein-Goodenough
```


Semantic similarity judgments: example



Semantic similarity judgments: results

Results on RG65 task:

- Padó and Lapata's (2007) dependency-based model: 0.62
- Dependency-based on Web corpus (Herdağdelen *et al.* 2009)
  - without SVD reduction: 0.69
  - with SVD reduction: 0.80
- Distributional memory (Baroni & Lenci 2010): 0.82
- Salient Semantic Analysis (Hassan & Mihalcea 2011): 0.86

And you?

```
> eval.similarity.correlation(RG65, M, convert=FALSE)
       rho  p.value missing     r r.lower r.upper
RG65 0.687 2.61e-10       0 0.678    0.52   0.791
> plot(eval.similarity.correlation( # cosine similarity
        RG65, M, convert=FALSE, details=TRUE))
```


Noun categorization

- In categorization tasks, subjects are typically asked to assign experimental items – objects, images, words – to a given category or group items belonging to the same category
  - categorization requires an understanding of the relationship between the items in a category
- Categorization is a basic cognitive operation presupposed by further semantic tasks
  - inference
    - ⋆ if X is a CAR then X is a VEHICLE
  - compositionality
    - ⋆ $\lambda y : \text{FOOD}\ \lambda x : \text{ANIMATE}\ \big[eat(x, y)\big]$
- "Chicken-and-egg" problem for relationship of categorization and similarity (cf. Goodman 1972, Medin et al. 1993)

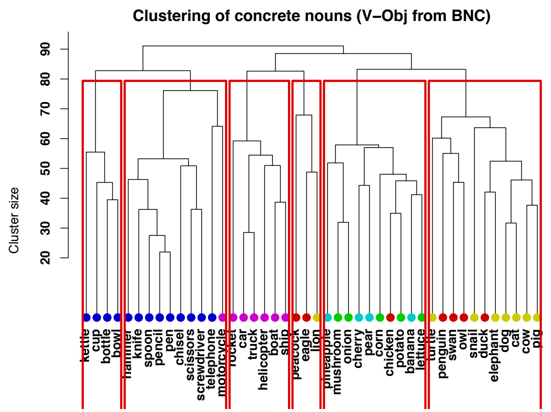
Noun categorization: the ESSLLI 2008 dataset

Dataset of 44 concrete nouns (ESSLLI 2008 Shared Task)

- 24 natural entities
  - 15 animals: 7 birds (*eagle*), 8 ground animals (*lion*)
  - 9 plants: 4 fruits (*banana*), 5 greens (*onion*)
- 20 artifacts
  - 13 tools (*hammer*), 7 vehicles (*car*)
- DSMs operationalize categorization as a clustering task
  1. for each noun w_i in the dataset, take its vector $\mathbf{w}_i$
  2. use a clustering method to group similar vectors $\mathbf{w}_i$
  3. evaluate whether clusters correspond to gold-standard semantic classes (purity, entropy, ...)

```
> ESSLLI08_Nouns[seq(1,40,5), ]
```


Noun categorization: example



- majority labels: tools, tools, vehicles, birds, greens, animals
- correct: 4/4, 9/10, 6/6, 2/3, 5/10, 7/11
- purity = 33 correct out of 44 = 75.0%

ESSLLI 2008 shared task

- Clustering experiments with CLUTO (Karypis 2003)
  - repeated bisection algorithm
  - 6-way (birds, ground animals, fruits, greens, tools and vehicles), 3-way (animals, plants and artifacts) and 2-way (natural and artificial entities) clusterings
- Quantitative evaluation
  - entropy – whether words from different classes are represented in the same cluster (best $= 0$)
  - purity – degree to which a cluster contains words from one class only (best $= 1$)
  - global score across the three clustering experiments

$$\sum_{i=1}^{3} \text{Purity}_i - \sum_{i=1}^{3} \text{Entropy}_i$$

ESSLLI 2008 shared task

model	*6-way*		*3-way*		*2-way*		*global*
	P	*E*	*P*	*E*	*P*	*E*	
Katrenko	89	13	100	0	80	59	197
Peirsman+	82	23	84	34	86	55	140
dep-typed (DM)	77	24	79	38	59	97	56
dep-filtered (DM)	80	28	75	51	61	95	42
window (DM)	75	27	68	51	68	89	44
Peirsman−	73	28	71	54	61	96	27
Shaoul	41	77	52	84	55	93	-106

`Katrenko`, `Peirsman+/-`, `Shaoul`: ESSLLI 2008 Shared Task
`DM`: Baroni & Lenci (2009)

And you?

```
> eval.clustering(ESSLLI08_Nouns, M)  # uses PAM clustering
```


Semantic priming

- Hearing/reading a "related" prime facilitates access to a target in various psycholing. tasks (naming, lexical decision, reading)
  - the word *pear* is recognized faster if heard/read after *apple*
- Hodgson (1991) single word lexical decision task, 136 prime-target pairs (cf. Padó & Lapata 2007)
  - similar amounts of priming found for different semantic relations between primes and targets ($\approx$ 23 pairs per relation)
    - ⋆ synonyms (synonym): *to dread/to fear*
    - ⋆ antonyms (antonym): *short/tall*
    - ⋆ coordinates (coord): *train/truck*
    - ⋆ super- and subordinate pairs (supersub): *container/bottle*
    - ⋆ free association pairs (freeass): *dove/peace*
    - ⋆ phrasal associates (phrasacc): *vacant/building*

Simulating semantic priming

McDonald & Brew (2004); Padó & Lapata (2007)

- DSMs and semantic priming
  1. for each related prime-target pair, measure cosine-based similarity between items (e.g., *to dread* / *to fear*)
  2. to estimate unrelated primes, take average of cosine-based similarity of target with other primes from same semantic relation (e.g., *to value* / *to fear*)
  3. similarity between related items should be significantly higher than average similarity between unrelated items
- Significant effects ($p < .01$) for all semantic relations
  - strongest effects for synonyms, antonyms & coordinates
- Alternative: classification task
  - given target and two primes, identify related prime (➜ multiple choice like TOEFL)

Analogy tasks

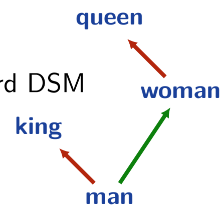
Mikolov *et al.* (2013b,a); Gladkova *et al.* (2016)

- Task: solve analogy problems such as
  - *man* : *woman* :: *king* : *queen*
  - *France* : *Paris* :: *Bulgaria* : *Sofia*
  - *learn* : *learned* :: *go* : *went*
  - *dog* : *animal* :: *strawberry* : *fruit*
- Approach 1: build DSM on word pairs as targets

$$\min_x d(\mathbf{v}_{\text{man:woman}}, \mathbf{v}_{\text{king:}x})$$

- Approach 2: use vector operations in single-word DSM

$$\mathbf{v}_{\text{queen}} \approx \mathbf{v}_{\text{king}} - \mathbf{v}_{\text{man}} + \mathbf{v}_{\text{woman}}$$

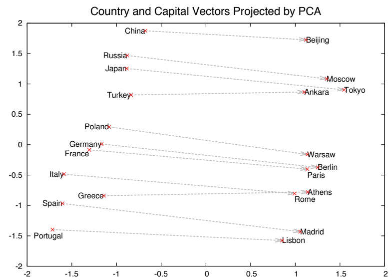


The Google analogy task

Mikolov *et al.* (2013b,a)

- Mikolov *et al.* (2013b,a) claim that their neural embeddings are good at solving analogy tasks

➥ Semantic features encoded in linear subdimensions



(Mikolov *et al.* 2013a, Fig. 2)

model	syntactic	semantic	
word2vec	64%	55%	(Mikolov *et al.* 2013b)
DSM	43%	60%	(Baroni *et al.* 2014)
FastText	82%	87%	(Mikolov *et al.* 2018)

The Google analogy task

Mikolov *et al.* (2013b,a)

But what is the task here?

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

(Mikolov *et al.* 2013b, Tab. 1)

The CogALex-V shared task

Santus *et al.* (2016)

Relation	Tag	Template	Example	Training	Testing
Synonymy	SYN	W2 can be used with the same meaning as W1	*candy-sweet, apartment-flat*	167	235
Antonymy	ANT	W2 can be used as the opposite of W1	*clean-dirty, add-take*	241	360
Hypernymy	HYPER	W1 is a kind of W2	*cannabis-plant, actress-human*	255	382
Part-whole meronymy	PART_OF	W1 is a part of W2	*calf-leg, aisle-store*	163	224
Random word	RANDOM	None of the above relations apply	*accident-fish, actor-mild*	2228	3059

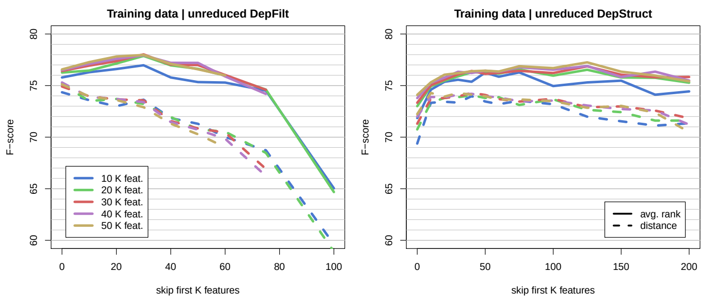
(Santus *et al.* 2016, Tab. 1)

- Task A: categorize pair as RANDOM vs. related (all other)
  - best system: $F_1 = 79.0\%$ (GHHH)
  - best DSM: $F_1 = 77.8\%$ (Mach5)
- Task B: categorize as SYN, ANT, HYPER, PART or RANDOM
  - best system: $F_1 = 44.5\%$ (LexNet)
  - best DSM: $F_1 = 29.5\%$ (Mach5 + SVM classifier)

Mach5: A cautionary tale

Evert (2016)

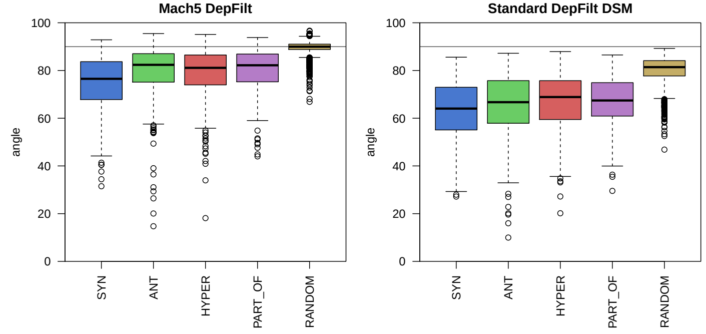
- DSM optimization for Task A was highly successful, but yielded counter-intuitive parameter settings
- ☞ Only feature ranks 20k–70k (DepFilt) / 50k–100k (DepStruct)



Mach5: A cautionary tale

Evert (2016)

- Nearest neighbours are unsatisfactory, e.g. for *play*: *playing* (54.1°), *star* (62.8°), *reunite* (62.9°), *co-star* (64.3°), *reprise* (64.4°), *player* (66.7°), *score* (68.5°), *audition* (69.2°), *sing* (69.4°), *actor* (69.5), *understudy* (69.6), *game* (70.3), ...



Outline

DSM evaluation in published studies

- One model, many tasks (Padó & Lapata 2007; Baroni & Lenci 2010; Pennington *et al.* 2014)
  - A novel DSM is proposed, with specific features & parameters
  - This DSM is tested on a range of different tasks (e.g. TOEFL, priming, semantic clustering)
- Incremental tuning of parameters (Bullinaria & Levy 2007, 2012; Kiela & Clark 2014; Polajnar & Clark 2014)
  - Several parameters (e.g., scoring measure, distance metric, dimensionality reduction)
  - Many tasks (e.g. TOEFL, semantic & syntactic clustering)
  - Varying granularity of parameter settings
  - One parameter (sometimes two) varied at a time, with all other parameters set to fixed values or optimized for each setting
  - Optimal parameter values are determined sequentially

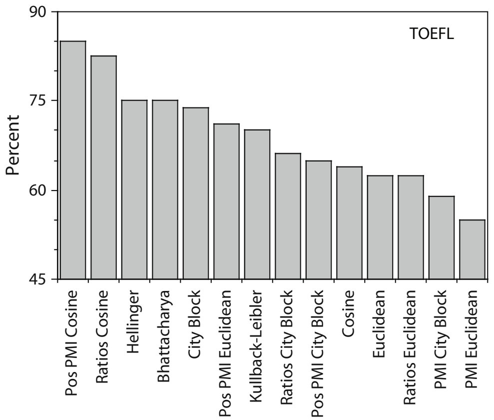
Outline

Bullinaria & Levy (2007, 2012)

- One of the first systematic evaluation studies
- Test influence of many standard parameter settings
  - frequency weighting + distance measure
  - co-occurrence window, structured vs. unstructured
  - corpus type & size, number of feature dimensions
  - dimensionality reduction (SVD), number of latent dimension
- In four different evaluation tasks
  - TOEFL
  - distance comparison: related word vs. 10 random words
  - semantic categorization: nearest-centroid classifier
  - syntactic categorization (2007)
  - semantic clustering of nouns (2012)
- Novel parameters
  - skipping of first latent dimensions (with highest variance)
  - Caron's (2001) P: power-scaling of singular values

TOEFL results: feature weighting & distance measure

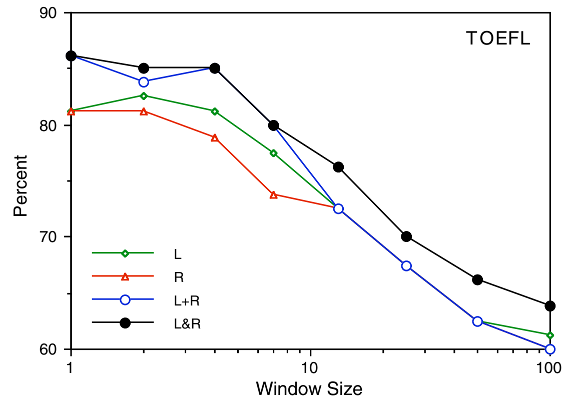
(Bullinaria & Levy 2007, p. 516, Fig. 1)



British National Corpus (BNC). Vectors not L2-normalized (frequency is L1-normalized). All other parameters optimized for each setting.

TOEFL results: size & type of co-occurrence window

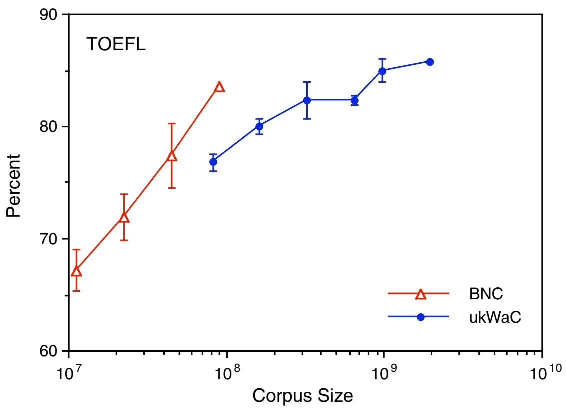
(Bullinaria & Levy 2012, p. 893, Fig. 1)



ukWaC Web corpus. Positive PMI + cosine (Bullinaria & Levy 2007). Number of feature dimensions optimized for each window size & task. No dimensionality reduction. L&R = structured surface context (left/right).

TOEFL results: corpus type & size

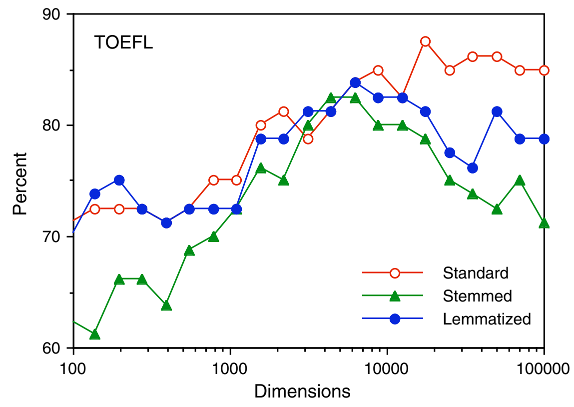
(Bullinaria & Levy 2012, p. 894, Fig. 2)



L+R context of size 1. Average + standard error over equally-sized corpus slices.
Other parameter settings unclear.

TOEFL results: feature dimensions & pre-processing

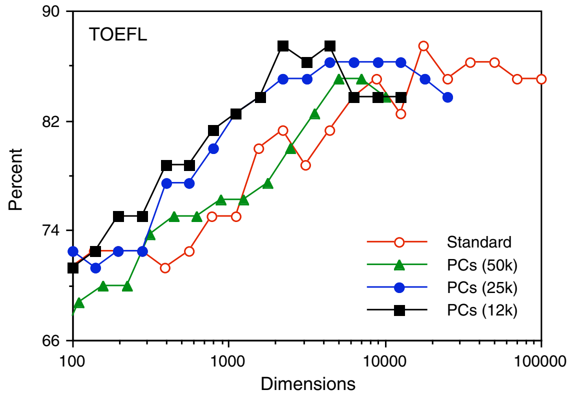
(Bullinaria & Levy 2012, p. 895, Fig. 4)



ukWaC corpus. L+R context of size 1. Other parameters presumably as above.

TOEFL results: dimensionality reduction

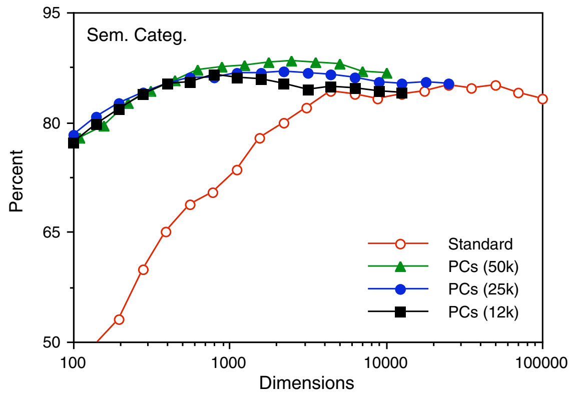
(Bullinaria & Levy 2012, p. 898, Fig. 5)



ukWaC corpus. Positive PMI + cosine. Standard = no dimensionality reduction.
Other: number of latent dimensions for 12k, 25k and 50k original feature dimensions.

Semantic categorization: dimensionality reduction

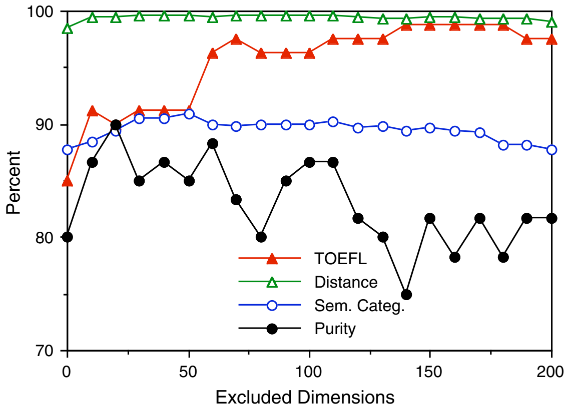
(Bullinaria & Levy 2012, p. 898, Fig. 5)



ukWaC corpus. Positive PMI + cosine. Standard = no dimensionality reduction.
Other: number of latent dimensions for 12k, 25k and 50k original feature dimensions.

Combined results: skipping first latent dimensions

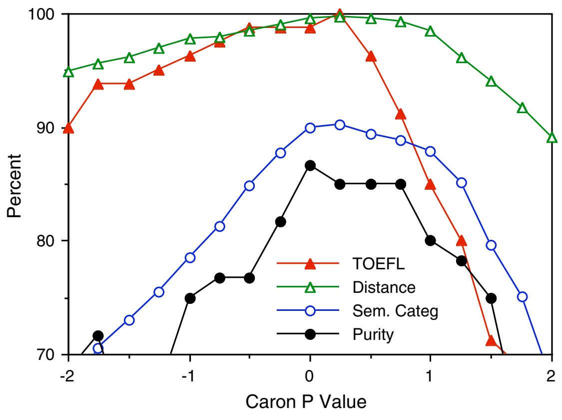
(Bullinaria & Levy 2012, p. 900, Fig. 7)



ukWaC corpus with standard settings. 50k feature dimensions reduced to 5000 latent dimensions.

TOEFL results: power scaling (Caron's P)

(Bullinaria & Levy 2012, p. 900, Fig. 7)



ukWaC corpus with standard settings. 50k feature dimensions reduced to 5000 latent dimensions. Neutral value is $P = 1$.

A (very) large-scale evaluation study

(Lapesa & Evert 2014)

Outline

Tasks

1. **Classification**
   - TOEFL80: multiple-choice classification task (Landauer & Dumais 1997)
2. **Correlation to Similarity Ratings**
   - RG65: 65 noun pairs (Rubenstein & Goodenough 1965)
   - WordSim353: 351 noun pairs (Finkelstein *et al.* 2002)
3. **Semantic Clustering**
   - Battig82: 82 nouns, 10 classes (Van Overschelde *et al.* 2004)
   - AP402: 402 nouns, 21 classes (Almuhareb 2006)
   - ESSLLI08_Nouns: 44 nouns, 6 classes
   - Mitchell: 60 nouns, 12 classes (Mitchell *et al.* 2008)

Distributional models: general features

- Term-term distributional semantic models (bag-of-words)
- Target terms (rows)
  - vocabulary from Distributional Memory (Baroni & Lenci 2010) + terms from evaluation datasets
  - 27522 lemma types
- Feature terms (columns)
  - filtered by part-of-speech (nouns, verbs, adjectives, adverbs)
  - further context selection determined by two model parameters

Distributional models were compiled and evaluated using the IMS Corpus Workbench[1], the UCS toolkit[2] and the wordspace package for R.

[1] http://cwb.sf.net/
[2] http://www.collocations.de/software.html

Evaluated parameters

Building the co-occurrence matrix

1. **Source corpus**: BNC, Wackypedia, UkWac

Our source corpora – standard choices in distributional semantics – differ in both size and quality. Is there a quantity/quality trade-off?

2. **Window** (= surface span)
   - **Direction**: directed (= structured), undirected
   - **Size**: 1, 2, 4, 8, 16 words

We expect those parameters to be crucial as they determine the granularity (direction) and amount (size) of shared context involved in the computation of similarity.

Evaluated parameters

Selecting dimensions from the co-occurrence matrix

3. **Feature selection**:
   - **Criterion**: frequency, number of non-zero entries
   - **Threshold**: top n dimensions (n = 5000, 10000, 20000, 50000, 100000)

How many context dimensions (words) are needed for DSMs to perform well in specific tasks? Are too many context dimensions detrimental? What is the best selection criterion?

Evaluated parameters

Weighting and scaling co-occurrence counts

4. **Feature scoring**: frequency, simple-ll, MI, Dice, t-score, z-score, tf.idf

Association measures represent an interpretation of co-occurrence frequency, and they emphasize different types of collocations (Evert 2008). Does this have an effect on DSM performance?

5. **Transformation**: none, logarithmic, square root, sigmoid

Transformations reduce the skewness of feature scores.

Evaluated parameters

Dimensionality reduction

6. **Dimensionality reduction** with randomized SVD:
   - **number of reduced dimensions**: 100, 300, 500, 700, 900
   - **number of skipped dimensions**: 0, 50, 100

Dimensionality reduction is expected to improve semantic representation and make computations more efficient. How does SVD interact with the other parameters? Bullinaria & Levy (2012) report improvements in some tasks (e.g. TOEFL) when the first latent dimensions (with highest variance) are discarded. Does this result generalize to our tasks/datasets?

Evaluated parameters

Computation and usage of distances

7. **Distance metric**: cosine (angular distance), manhattan

Both are symmetric, while cognitive processes are often asymmetric

8. **Index of distributional relatedness**
   - **distance**: $\mathrm{dist}(a, b)$
   - **neighbor rank**, calculated differently for different tasks:
     - TOEFL: backward rank, i.e. $\mathrm{rank}(b, a)$
     - Ratings and Clustering: average of logarithmic forward and backward rank, i.e. $\big(\log \mathrm{rank}(a, b) + \log \mathrm{rank}(b, a)\big)/2$

This parameter allows us to account for asymmetries: $\mathrm{rank}(b, a)$ is different from $\mathrm{rank}(a, b)$. While cognitively plausible, neighbor rank is computationally expensive: does it improve the performance of DSMs?

Outline

How many models did we end up with?

... and how do we make sense of all those results?

- We tested all the possible parameter combinations (we will see later that this is crucial for our evaluation methodology)
- 537600 model runs (33600 in the unreduced setting, 504000 in the reduced setting)
- The models were generated and evaluated on a large HPC cluster at FAU Erlangen-Nürnberg as well as servers at the University of Stuttgart, within approximately 5 weeks

Evaluation methodology: linear regression

Our proposal for a robust evaluation of DSM parameters

- Attempts to predict the values of a "dependent" variable from one or more "independent" variables and their combinations
- Is used to understand which independent variables are closely related to the dependent variable, and to explore the forms of these relationships

Example

Dependent variable: income
Independent variables: gender, age, ethnicity, education level, first letter of the surname (hopefully not significant)

Evaluation methodology: linear regression

Our proposal for a robust evaluation of DSM parameters

We use linear models to analyze the influence of different DSM parameters and their combinations on DSM performance

- dependent variable = performance
  (accuracy, correlation coefficient, purity)
- independent variables = model parameters
  (e.g., source corpus, window size, window direction)

We want to understand which of the parameters are related to the dependent variable, i.e., we want to find the parameters whose manipulation has the strongest effect on DSM performance.

DSM evaluation and linear regression

Toy example: a $2 \times 2 \times 2$ design

Corpus	Window size	Window direction	Accuracy
ukWaC	1	directed	88
ukWaC	16	undirected	92
ukWaC	1	directed	91
ukWaC	16	undirected	93
BNC	1	undirected	80
BNC	16	undirected	53
BNC	1	directed	72
BNC	16	directed	71

$$\begin{aligned} \mathit{Accuracy} = {} & \beta_0 + \beta_1(\text{corpus}) + \beta_2(\text{window size}) + \beta_3(\text{window direction}) \\ & + \beta_4(\text{corpus: window size}) + \beta_5(\text{corpus: window direction}) + \\ & + \beta_6(\text{window size: window direction}) + \epsilon \end{aligned}$$

*we're aware that this regression model is almost saturated ...

DSM evaluation and linear regression

Analysis of variance

Goal: quantify the impact of a specific parameter (or interaction) on DSM performance, in terms of the proportion of variance explained by the parameter

Key notions:

- R^2 (R squared)
  - proportion of explained variance, i.e.

$$1 - \frac{\text{residual variance of } \epsilon}{\text{variance of dependent variable}}$$

  - calculated (i) for the full model (➜ how well the model exlains the experimental results) as well as (ii) for specific parameters and interactions (quantifying how much they contribute to predictions)
- Feature ablation

DSM evaluation and linear regression

Analysis of variance: feature ablation

Feature ablation

Proportion of variance explained by a parameter together with all its interactions, corresponding to the reduction in R^2 of the linear model fit if this parameter is left out.

In our toy model with 3 parameters and all two-way interactions:

- Ablation(corpus) = R^2(corpus) + R^2(corpus: window size) + R^2(corpus: window direction)
- Ablation(window size) = R^2(window size) + R^2(corpus: window size) + R^2(window size: window direction)
- Ablation(window direction) = R^2(window direction) + R^2(corpus: window direction) + R^2(window size: window direction)

Outline

TOEFL multiple-choice classification task

Introducing the task

A collection of 80 multiple-choice questions from a synonym task in the Test Of English as a Foreign Language (TOEFL)

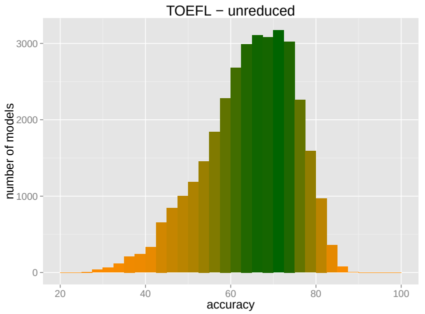
TOEFL dataset

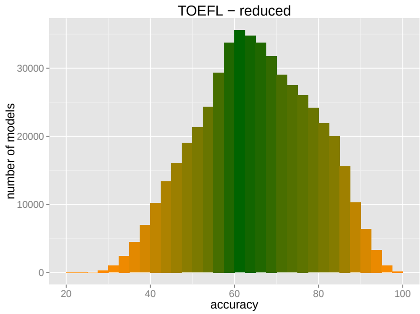
Target: *consume* – Choices: *breed*, *catch*, *eat*, *supply*
Target: *constant* – Choices: *accidental*, *continuing*, *instant*, *rapid*
Target: *concise* – Choices: *free*, *positive*, *powerful*, *succinct*

- A **classification** task
- If DSMs capture synonymy relations, we expect that the distance between the target and the correct choice will be smaller than to the wrong choices
- Performance: **% accuracy**

TOEFL task: performance

Unreduced versus Reduced Experiments



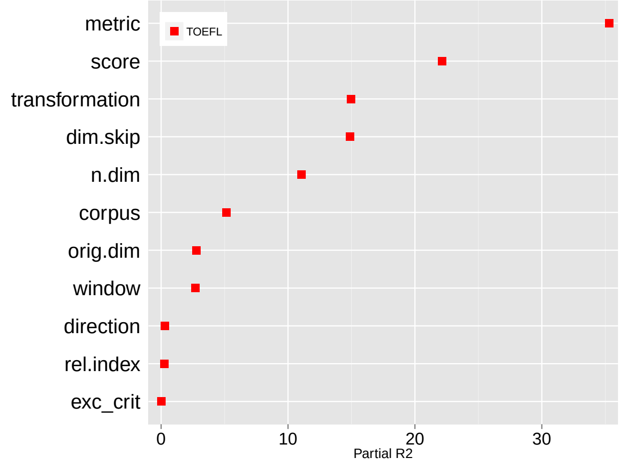


Min: 25 ; Max: 87.5 ; Mean: 63.9

Min: 18.7; Max: 97.4; Mean: 64.4

TOEFL task: parameters and explained variance

Reduced setting: feature Ablation (model R^2: 89%)



TOEFL task: interactions

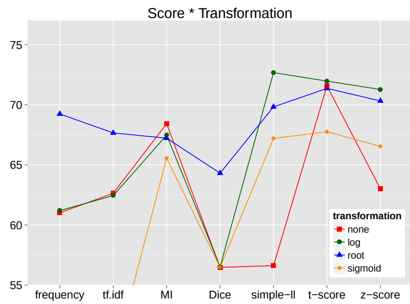
Reduced setting ($R^2 > 0.5$)

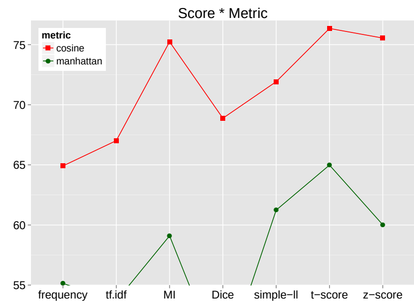
Interaction	Df	R^2
score:transformation	18	7.42
metric:dim.skip	2	4.44
score:metric	6	1.77
metric:orig.dim	4	0.98
window:transformation	12	0.91
corpus:score	12	0.84
score:orig.dim	24	0.64
metric:n.dim	4	0.63

TOEFL task: interactions, R^2

TOEFL task: Metric, Score, Transformation

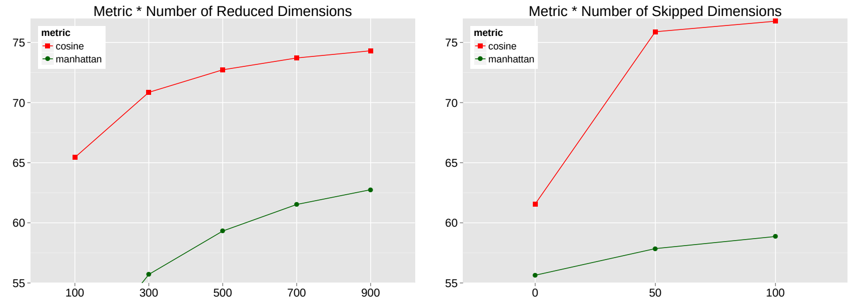
Partial effect displays (Fox 2003)





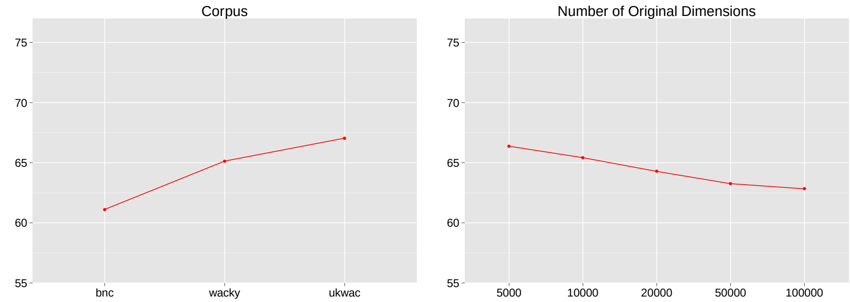
TOEFL task: Dimensionality Reduction

Partial effect displays (Fox 2003)



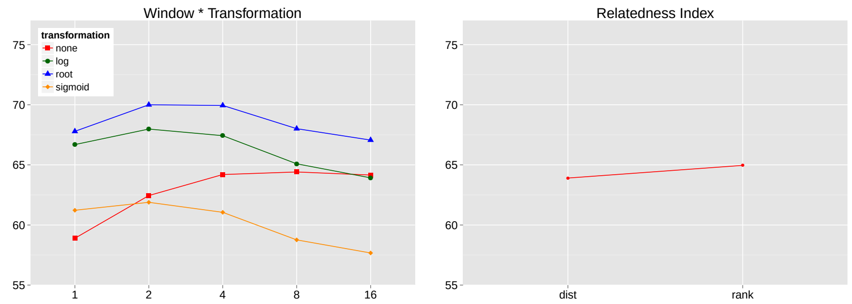
TOEFL task: Corpus and Number of Feature Dimensions

Partial effect displays (Fox 2003)



TOEFL task: Window and Relatedness Index

Partial effect displays (Fox 2003)



TOEFL task: summary

TOEFL: best setting

- Corpus: ukWac
- Window: undirected, 2 words
- Feature selection: top 5000/10000 dimensions, based on frequency
- Score * Transformation: simple-ll * log
- Dimensionality Reduction: 900 latent dimensions, skipping the first 100
- Distance Metric: cosine
- Index of Distributional Relatedness: neighbor rank

DSMs and similarity ratings

Introducing the task

RG65

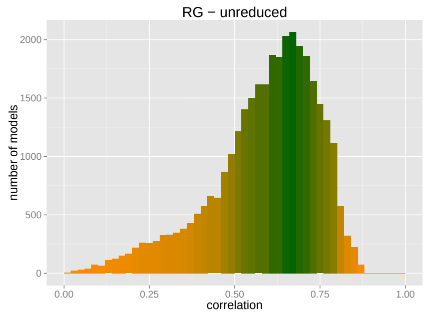
65 pairs, rated from 0 to 4
gem – jewel: 3.94
grin – smile: 3.46
fruit – furnace: 0.05

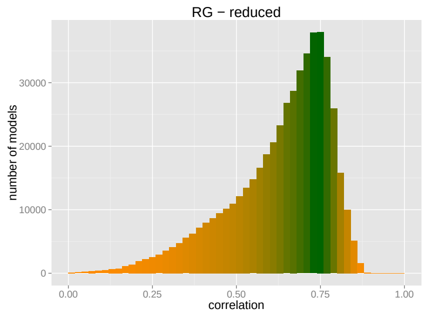
WordSim353

353 pairs, rated from 1 to 10
announcement – news: 7.56
weapon – secret: 6.06
travel – activity: 5.00

- A **prediction** task
- If distributional representation are close to speakers' conceptual representations, we expect to find some **correlation** between distance in the semantic space and speaker's judgments concerning semantic similarity
- Performance: **Pearson correlation** *r*

Similarity ratings: performance on RG65

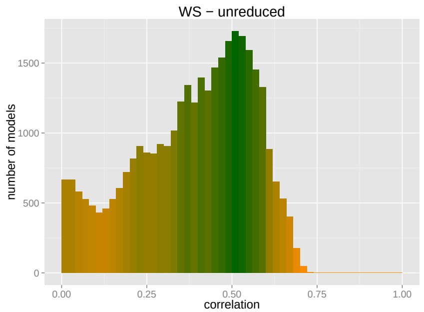


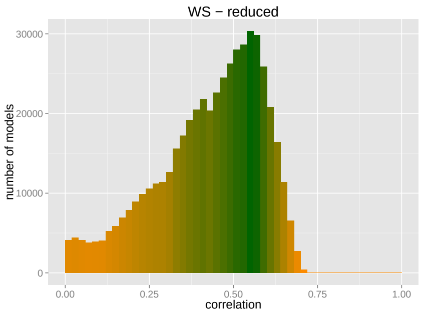


Min: 0.01; Max: 0.88; Mean 0.59
unreduced

Min: 0.00; Max: 0.89; Mean: 0.63
reduced

Similarity ratings: performance on WordSim353



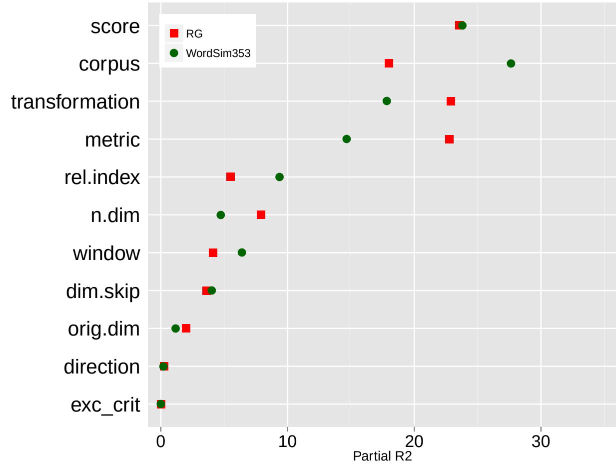


Min: 0.00; Max: 0.73; Mean: 0.39
unreduced

Min: 0.00; Max: 0.73; Mean: 0.43
reduced

Similarity ratings: parameters and explained variance

Reduced setting: feature ablation (full model R^2: RG65 86%; WS353 90%)



Similarity ratings: interactions

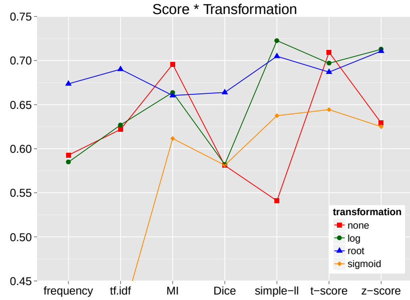
Reduced setting ($R^2 > 0.5$)

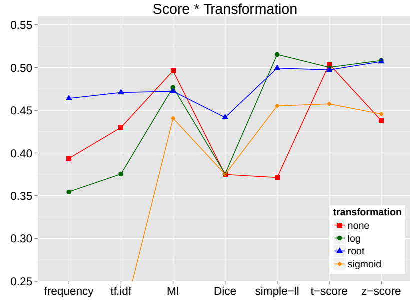
Interaction	Df	RG65	WordSim353
score:transf	18	10.28	8.66
metric:n.dim	4	2.18	1.42
window:transf	12	1.43	1.01
corpus:metric	2	1.83	0.51
score:metric	6	1.91	0.59
metric:orig.dim	4	1.08	0.62
corpus:score	12	0.77	0.82
window:score	24	0.77	0.69
score:dim.skip	12	0.58	0.85

Similarity ratings: interactions, R^2

Similarity ratings: Score, Transformation

Partial effect displays (Fox 2003)

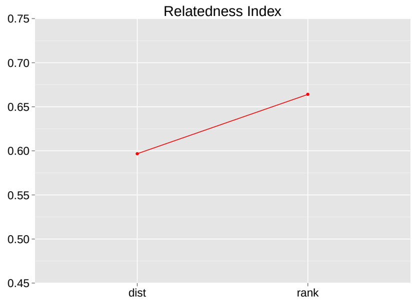


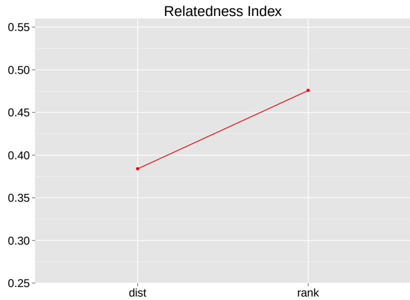


Rubenstein & Goodenough WordSim-353

Similarity ratings: Relatedness Index

Partial effect displays (Fox 2003)





Rubenstein & Goodenough WordSim-353

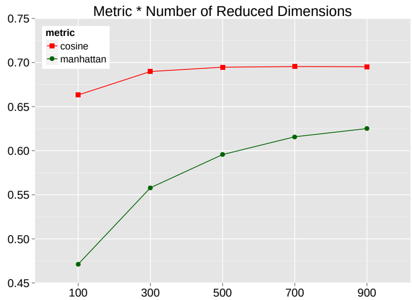
Similarity ratings: Metric, Number of Feature Dimensions

Partial effect displays (Fox 2003)

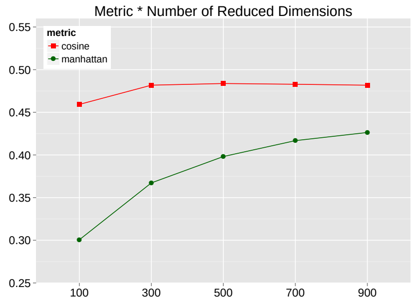
Rubenstein & Goodenough WordSim-353

Similarity ratings: Number of Latent Dimensions

Partial effect displays (Fox 2003)



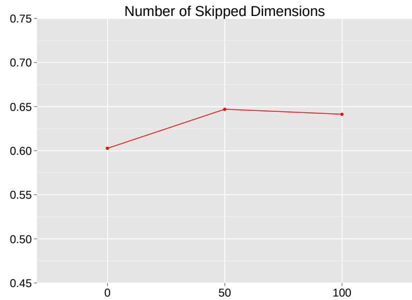
Rubenstein & Goodenough



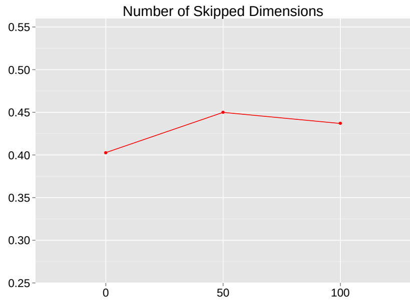
WordSim-353

Similarity ratings: Number of Skipped Dimensions

Partial effect displays (Fox 2003)



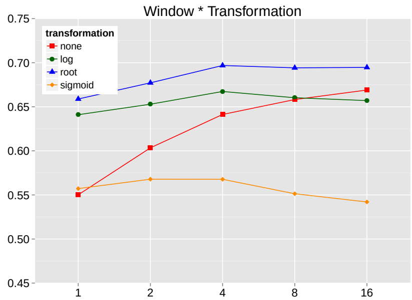
Rubenstein & Goodenough



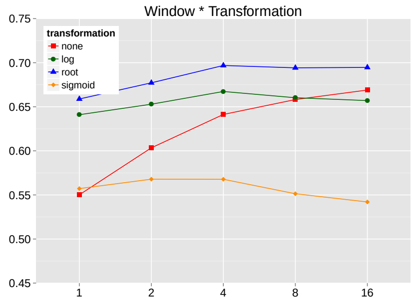
WordSim-353

Similarity ratings: Window Size, Transformation

Partial effect displays (Fox 2003)



Rubenstein & Goodenough



WordSim-353

Summing up: Ratings

Ratings: best setting

- Corpus: wacky
- Window: undirected, 4 words
- Feature selection: top 20000/50000 dimensions, based on frequency
- Score * Transformation: simple-ll * log
- Dimensionality Reduction: 300 latent dimensions, skipping the first 50
- Distance Metric: cosine
- Index of Distributional Relatedness: neighbor rank

DSMs and semantic clustering

Introducing the task

Almuhareb & Poesio

402 nouns, 21 classes

day ⟹ TIME
kiwi ⟹ FRUIT
kitten ⟹ ANIMAL
volleyball ⟹ GAME

BATTIG set

82 nouns, 10 classes

chicken ⟹ BIRD
bear ⟹ LAND_MAMMAL
pot ⟹ KITCHENWARE
oak ⟹ TREE

ESSLLI categorization task

44 nouns, 6 classes

potato ⟹ GREEN
hammer ⟹ TOOL
car ⟹ VEHICLE
peacock ⟹ BIRD

MITCHELL set

60 nouns, 12 classes

ant ⟹ INSECT
carrot ⟹ VEGETABLE
train ⟹ VEHICLE
cat ⟹ ANIMAL

DSMs and semantic clustering

Introducing the task

- A **categorization** task
- If distributional representations approximate human conceptual representations, we expect word categorization based on distributional features to produce concept clusters similar to those in the gold standard datasets
- Performance: **cluster purity**
  - classification accuracy for optimal cluster labelling
  - percentage of nouns that belong to the majority category within their cluster
- Partitioning around medoids (Kaufman & Rousseeuw 1990)
  - implemented as `pam()` in R standard library
  - direct comparison ➜ equal to or even better than CLUTO
  - works with arbitrary dissimilarity matrix

Semantic clustering: performance

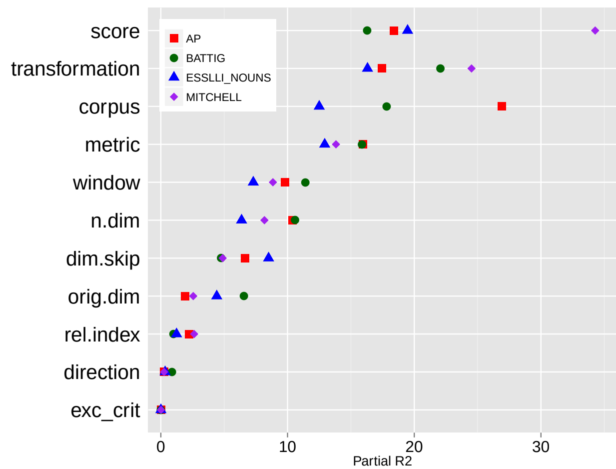
Overview: unreduced versus reduced experiments

Dataset	Unreduced			Reduced		
	Min	Max	Mean	Min	Max	Mean
AP	0.15	0.73	0.56	0.13	0.76	0.54
BATTIG	0.28	0.99	0.77	0.23	0.99	0.78
ESSLLI	0.32	0.93	0.72	0.32	0.98	0.72
MITCHELL	0.26	0.97	0.68	0.27	0.97	0.69

Semantic clustering: summary of performance (purity)

Semantic clustering: parameters and explained variance

Feature ablation (model R^2 – AP: 82%; BATTIG: 77%; ESSLLI 58%; MITCHELL 73%)



Semantic clustering: interactions

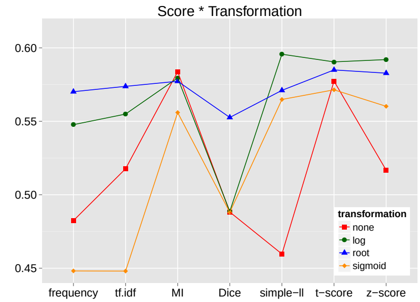
Reduced setting ($R^2 > 0.5$)

Interaction	Df	AP	BATTIG	ESSLLI	MITCHELL
score:transformation	18	7.10	7.95	7.56	11.42
window:metric	4	2.22	1.26	2.97	2.72
metric:n.dim	4	3.29	3.16	2.03	0.58
metric:dim.skip	2	2.25	1.54	2.77	0.86
window:transformation	12	2.00	2.95	0.88	2.66
corpus:metric	2	1.42	2.91	2.79	1.11
corpus:window	8	2.36	1.18	1.49	1.23
score:dim.skip	12	0.56	1.15	0.99	1.39
window:score	24	0.74	0.77	0.54	0.65

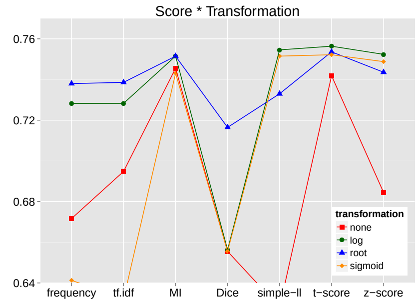
Clustering datasets: interactions, R^2

Semantic clustering: Score, Transformation

Partial effect displays (Fox 2003)



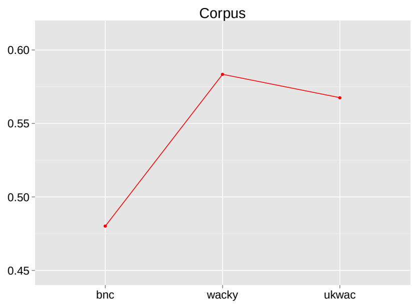
Almuhareb & Poesio



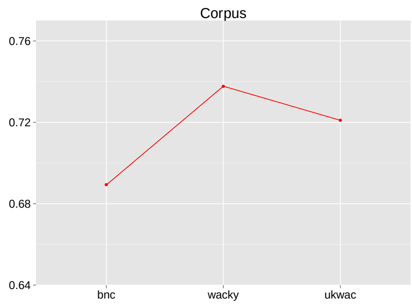
ESSLLI 2008

Semantic clustering: Corpus

Partial effect displays (Fox 2003)



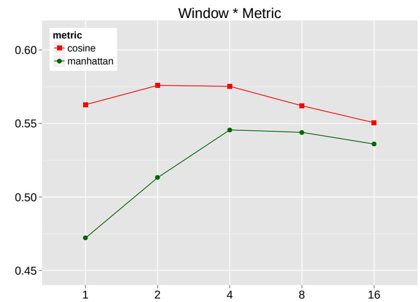
Almuhareb & Poesio



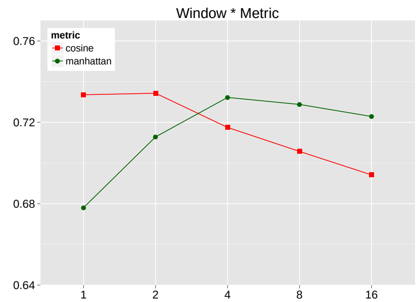
ESSLLI 2008

Semantic clustering: Window Size, Metric

Partial effect displays (Fox 2003)



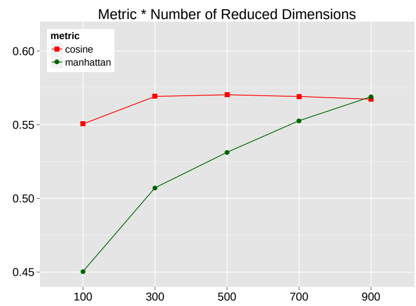
Almuhareb & Poesio



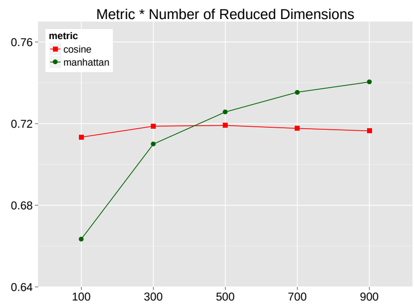
ESSLLI 2008

Semantic clustering: Metric, Number of Latent Dimensions

Partial effect displays (Fox 2003)



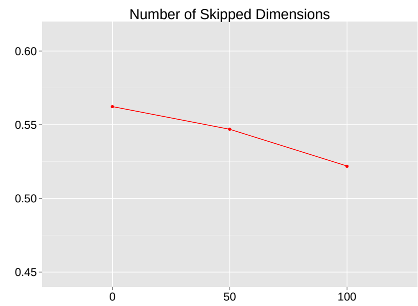
Almuhareb & Poesio



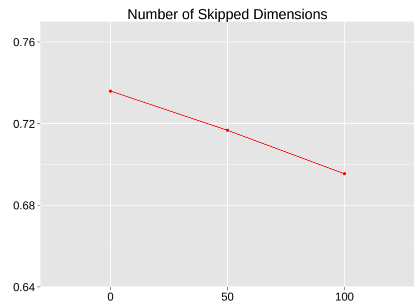
ESSLLI 2008

Semantic clustering: Number of Skipped Dimensions

Partial effect displays (Fox 2003)



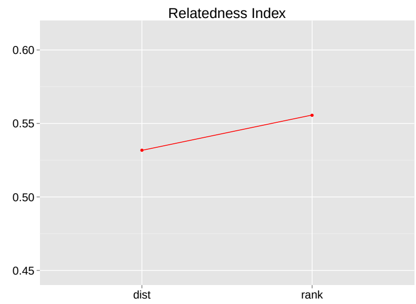
Almuhareb & Poesio



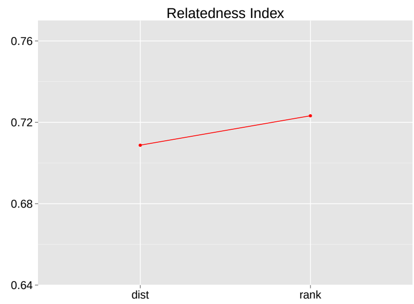
ESSLLI 2008

Semantic clustering: Relatedness Index

Partial effect displays (Fox 2003)



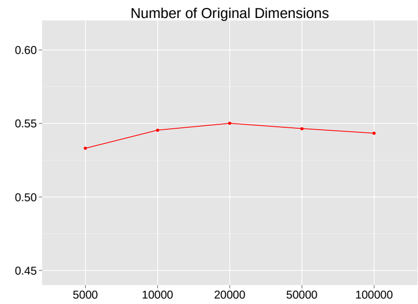
Almuhareb & Poesio



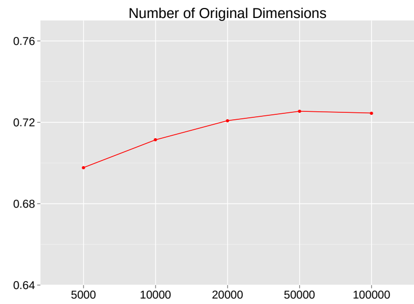
ESSLLI 2008

Semantic clustering: Number of Feature Dimensions

Partial effect displays (Fox 2003)



Almuhareb & Poesio



ESSLLI 2008

Summing up: Semantic Clustering

Clustering: best setting

- Corpus: wacky
- Window: undirected, 4 words
- Feature selection: top 50000 dimensions, based on frequency
- Score * Transformation: simple-ll * log (or t-score * log)
- Dimensionality Reduction: 300/500 latent dimensions, no skipping necessary
- Distance Metric: cosine
- Index of Distributional Relatedness: neighbor rank

Outline

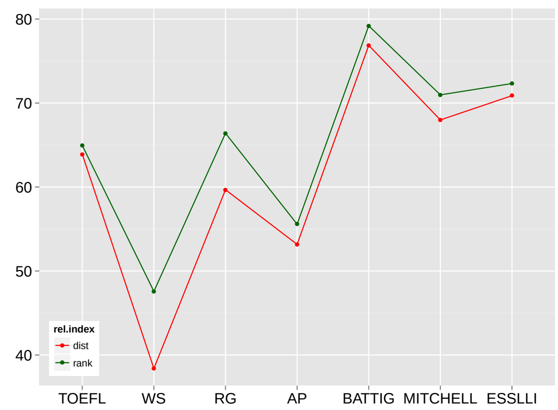
Does our evaluation methodology work?

1. What are the most explanatory parameters?
2. By inspecting the effect plots, we identified best settings for every dataset: what is the performance of such best settings? Are they close to the best runs in the experiment?
3. Is it possible to identify a general best setting that performs reasonably well across all tasks?

Summary: parameters

- Parameters with strong effect on DSM performance and homogeneous behavior across tasks and datasets
  - score
  - transformation
  - distance metric
- Parameters with strong effect on DSM performance, but differences across tasks
  - dimensionality reduction parameters
  - window
  - corpus (to a lesser extent)
- A less crucial parameter with homogeneous behavior
  - number of context dimensions
- Parameters that have no or little effect on DSM performance
  - criterion for context selection
  - direction of the context window

How about the index of distributional relatedness?



Best settings and their performance

dataset	corpus	w	o.dim	sc	tr	m	rel.ind	n.dim	d.sk	best.s	best.m
TOEFL	ukwac	2	5k	s-ll	log	cos	rank	900	100	92.5	98.75
WS	wacky	4	50k	s-ll	log	cos	rank	300	50	0.67	0.73
RG	wacky	4	50k	s-ll	log	cos	rank	300	50	0.86	0.89
AP	wacky	4	20k	s-ll	log	cos	rank	300	0	0.69	0.76
BATTIG	wacky	8	50k	s-ll	log	cos	rank	500	0	0.98	0.99
ESSLLI	wacky	2	20k	t-sc	log	cos	rank	300	0	*0.77*	*0.98*
MITCHELL	wacky	4	50k	s-ll	log	cos	rank	500	0	0.88	0.97

Best settings for each dataset

w = window size, o.dim = number of feature dimensions, sc = scoring function, tr = transformation, m = metric, d.sk = number of skipped dimensions, best.s = performance of best setting for this dataset, best.m = performance of best run for this dataset

General settings

task	corpus	w	o.dim	sc	tr.	m	rel.ind	n.dim	d.sk
TOEFL	ukwac	2	5k	s-ll	log	cos	rank	900	100
Rating	wacky	4	50k	s-ll	log	cos	rank	300	50
Clustering	wacky	4	50k	s-ll	log	cos	rank	500	0
General	wacky	4	50k	s-ll	log	cos	rank	500	50

General best settings

Task	TOEFL	RATINGS	CLUSTERING	GENERAL	SoA
TOEFL	92.5	85.0	75.0	90.0	100.0
RG	0.85	0.86	0.84	0.87	0.86
WS	0.60	0.67	0.64	0.68	0.81
AP402	0.60	0.66	0.67	0.67	0.79
BATTIG	0.85	0.91	0.98	0.90	0.96
ESSLLI	0.70	0.77	0.80	0.77	0.91
MITCHELL	0.73	0.83	0.88	0.83	0.94

General best settings – Performance

Conclusion

- Our results show that it is possible to find a single DSM configuration that performs relatively well on every task
- The most explanatory parameters show similar behavior across all tasks and datasets
  - Simple-ll * Logarithmic Transformation
  - Cosine Distance
- Parameters that show variation determine the amount and nature of the shared context
  - Context window: 4 is a good compromise solution
  - Dimensionality reduction: skipping the first dimensions (but not too many) generally helps
  - Number of Feature Terms (to a lesser extent)

Conclusion

- Among the source corpora, WaCkypedia appears to be a better option than UkWaC for all tasks but TOEFL
  - A good trade-off between quantity and quality?
- As an index of distributional relatedness, neighbor rank is always better than distance, even if its contribution to model performance varies across tasks
  - Perhaps some tasks/datasets are less asymmetric than others?
  - may need to exploit directionality in a more granular way
- But remember the Mach5 lession:
  good evaluation results $\neq$ accurate semantic representation

References I

Almuhareb, Abdulrahman (2006). *Attributes in Lexical Acquisition*. Ph.D. thesis, University of Essex.

Baroni, Marco and Lenci, Alessandro (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, **20**(1).

Baroni, Marco and Lenci, Alessandro (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 673–712.

Baroni, Marco and Lenci, Alessandro (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK.

Baroni, Marco; Dinu, Georgiana; Kruszewski, Germán (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, Baltimore, MD.

Bruni, Elia; Tran, Nam Khanh; Baroni, Marco (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, **49**, 1–47.

Budanitsky, Alexander and Hirst, Graeme (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, **32**(1), 13–47.

References II

Bullinaria, John A. and Levy, Joseph P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, **39**(3), 510–526.

Bullinaria, John A. and Levy, Joseph P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, **44**(3), 890–907.

Caron, John (2001). Experiments with LSA scoring: Optimal rank and basis. In M. W. Berry (ed.), *Computational Information Retrieval*, pages 157–169. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Evert, Stefan (2008). Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin, New York.

Evert, Stefan (2016). CogALex-V shared task: Mach5 – a traditional DSM approach to semantic relatedness. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*, pages 92–97, Osaka, Japan.

Finkelstein, Lev; Gabrilovich, Evgeniy; Matias, Yossi; Rivlin, Ehud; Solan, Zach; Wolfman, Gadi; Ruppin, Eytan (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, **20**(1), 116–131.

References III

Fox, John (2003). Effect displays in r for generalised linear models. *Journal of Statistical Software*, **8**(15), 1–27.

Gladkova, Anna; Drozd, Aleksandr; Matsuoka, Satoshi (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California.

Hassan, Samer and Mihalcea, Rada (2011). Semantic relatedness using salient semantic analysis. In *Proceedings of the Twenty-fifth AAAI Conference on Artificial Intelligence*.

Herdağdelen, Amaç; Erk, Katrin; Baroni, Marco (2009). Measuring semantic relatedness with vector space models and random walks. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 50–53, Suntec, Singapore.

Hill, Felix; Reichart, Roi; Korhonen, Anna (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, **41**(4), 665–695.

Hodgson, James M. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes*, **6**(3), 169–205.

References IV

Hutchison, Keith A.; Balota, David A.; Neely, James H.; Cortese, Michael J.; Cohen-Shikora, Emily R.; Tse, Chi-Shing; Yap, Melvin J.; Bengson, Jesse J.; Niemeyer, Dale; Buchanan, Erin (2013). The semantic priming project. *Behavior Research Methods*, **45**(4), 1099–1114.

Kaufman, Leonard and Rousseeuw, Peter J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.

Kiela, Douwe and Clark, Stephen (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 21–30, Gothenburg, Sweden.

Landauer, Thomas K. and Dumais, Susan T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, **104**(2), 211–240.

Lapesa, Gabriella and Evert, Stefan (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, **2**, 531–545.

McDonald, Scott and Brew, Chris (2004). A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL '04)*, pages 17–24, Barcelona, Spain.

References V

Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S.; Dean, Jeff (2013a). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119. Curran Associates, Inc.

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey (2013b). Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations 2013*.

Mikolov, Tomas; Grave, Edouard; Bojanowski, Piotr; Puhrsch, Christian; Joulin, Armand (2018). Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 52–55, Miyazaki, Japan.

Mitchell, Tom M.; Shinkareva, Svetlana V.; Carlson, Andrew; Chang, Kai-Min; Malave, Vicente L.; Mason, Robert A.; Just, Marcel Adam (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, **320**, 1191–1195.

Padó, Sebastian and Lapata, Mirella (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, **33**(2), 161–199.

Pennington, Jeffrey; Socher, Richard; Manning, Christopher D. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP 2014*.

References VI

Polajnar, Tamara and Clark, Stephen (2014). Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 230–238, Gothenburg, Sweden.

Rapp, Reinhard (2004). A freely available automatically generated thesaurus of related words. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 395–398.

Rogers, Anna; Hosur Ananthakrishna, Shashwath; Rumshisky, Anna (2018). What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 2690–2703, Santa Fe, NM. Association for Computational Linguistics.

Rubenstein, Herbert and Goodenough, John B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, **8**(10), 627–633.

Santus, Enrico; Gladkova, Anna; Evert, Stefan; Lenci, Alessandro (2016). The CogALex-V shared task on the corpus-based identification of semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*, pages 69–79, Osaka, Japan.

Turney, Peter D. (2006). Similarity of semantic relations. *Computational Linguistics*, **32**(3), 379–416.

References VII

Van Overschelde, James; Rawson, Katherine; Dunlosky, John (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, **50**, 289–335.