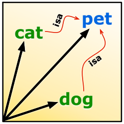
Distributional Semantic Models

Part 3: Evaluation – is my DSM "good"?
Part 4: DS beyond NLP: Linguistic Issues

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

[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



The problem

"The distributional hypothesis, as motivated by the works of Zellig Harris, is a strong methodological claim with a weak semantic foundation. It states that differences of meaning correlate with differences of distribution,but it neither specifies what kind of distributional information we should look for, nor what kind of meaning differences it mediates." (Sahlgren 2008)

The solution

Which kind of meaning nuance is my DSM capturing (if any)?

1. Parameter manipulation
   - ... what kind of information should we look for?
   - ... after yesterday's lecture, we are all experts and we know how many different options we have!

2. **Evaluation: { tasks + datasets }**
   - ... what kind of meaning differences are we capturing?
   - ... in a way, while we extract/visualize neighbors (task) our intuition about "what a good neighbor is" is the dataset

3. **Interpretation** of the evaluation results
   - crucial issue, often disregarded or oversimplified

Outline

Tasks & Datasets

- **Tasks** are experimental setups to test DSM representations:
  - **Classification (multiple choice)**: given a target word, pick the "best" from a set of candidates (whatever best means)
  - **Correlation**: do DSM similarities approximate values which quantify semantic simliarity/relatedness (ratings, reaction times)?
  - **Categorization**: do DSM similarities group words in a "reasonable" way?
- **Datasets** are the external "ground truth" and contribute the semantic "nuance" to the evaluation
  - Collected ad-hoc for DSM evaluation or (often) existing independently of it
    - e.g., TOEFL, similarity ratings, experimental items from psycholinguistic experiments)

{Task + Dataset} as operationalization of a hypothesis, e.g..
DSM similarity as synonymy → multiple choice task + TOEFL

Tasks

Instrinsic vs. Extrinsic tasks

- **Intrinsic evaluation** the semantic representations produced by the DSM are evaluated *directly*
  - The DSM is the *only* responsible for the performance
- **Extrinsic evaluation**: the DSM representations are input to further tasks, whose performance is then evaluated, e.g.,
  - DSM vectors as input of a machine learning classifier → accuracy of the classifier
  - DSM vectors to improve a machine translation system → BLEU score of the MT

Datasets

Reminder: the many facets of DSM similarity

- 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*

Datasets for intrinsic evaluation of attributional similarity/relatedness

- 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 (intrinsic & extrinsic, ML)
  - 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
```

Outline

The TOEFL synonym task

- The TOEFL dataset (80 items)
  - Target: *show*
    Candidates: *demonstrate*, *publish*, *repeat*, *postpone*
  - Target *costly*
    Candidates: *beautiful*, *complicated*, *expensive*, *popular*

- DSMs and TOEFL
  1. take vectors of the target ($\mathbf{t}$) and of the candidates ($\mathbf{c}_1 \ldots \mathbf{c}_n$)
  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 $\mathbf{t}$

```
> 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)
```

Outline

Semantic similarity judgments

RG65

65 pairs, rated from 0 to 4

gem – jewel: 3.94
grin – smile: 3.46
fruit – furnace: 0.05

WordSim353

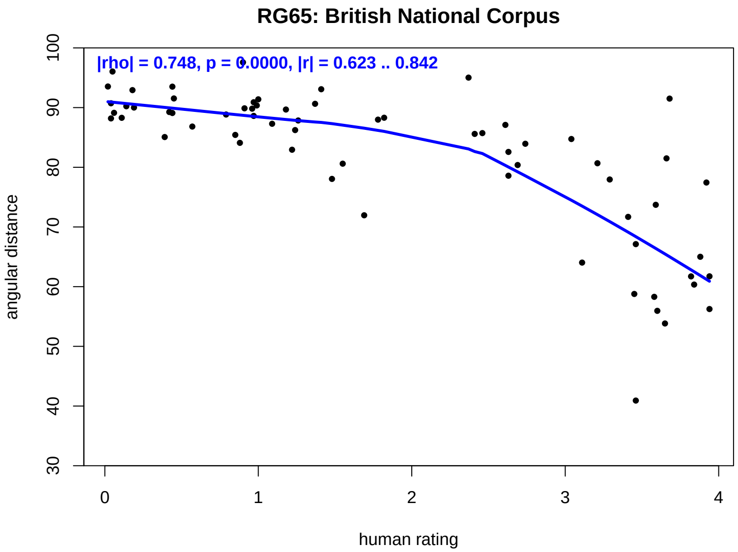
353 pairs, rated from 1 to 10

announcement – news: 7.56
weapon – secret: 6.06
travel – activity: 5.00

- DSMs vs. Ratings: operationalization
  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 correlation between vector distances and R&G average judgments (Padó & Lapata 2007)

```
> RG65[seq(0,65,5), ]
> head(WordSim353)
```

Semantic similarity judgments: example



Semantic similarity judgments: results

Results on RG65 task (Pearson):

- 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))
```

Outline

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
    - $\lambda y : \text{FOOD}\ \lambda x : \text{ANIMATE}\ \left[eat(x, y) \right]$
- "Chicken-and-egg" problem for relationship of categorization and similarity (cf. Goodman 1972, Medin et al. 1993)

Noun categorization: datasets

ESSLLI08 (on focus today)

44 nouns, 6 classes

potato ⟹ GREEN

hammer ⟹ TOOL

car ⟹ VEHICLE

peacock ⟹ BIRD

BATTIG set

82 nouns, 10 classes

chicken ⟹ BIRD

bear ⟹ LAND_MAMMAL

pot ⟹ KITCHENWARE

oak ⟹ TREE

Almuhareb & Poesio

402 nouns, 21 classes

day ⟹ TIME

kiwi ⟹ FRUIT

kitten ⟹ ANIMAL

volleyball ⟹ GAME

MITCHELL set

60 nouns, 12 classes

ant ⟹ INSECT

carrot ⟹ VEGETABLE

train ⟹ VEHICLE

cat ⟹ ANIMAL

Noun categorization: the ESSLLI 2008 dataset

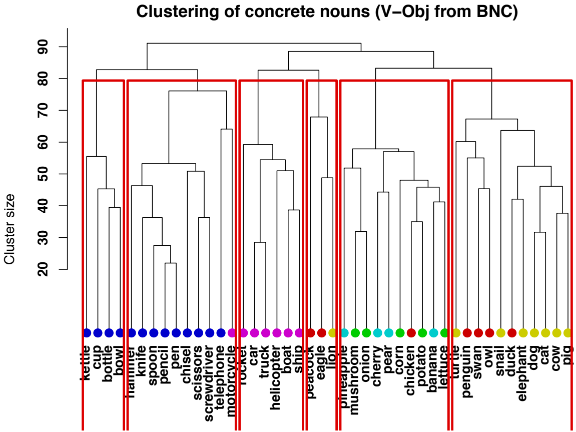
Dataset of 44 concrete nouns (ESSLLI 2008 Shared Task)

- 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 operationalizes 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

- Experiments:
  - 6-way (birds, ground animals, fruits, greens, tools and vehicles), 3-way (animals, plants and artifacts) and 2-way (natural and artificial entities) clusterings
- Evaluation scores:
  - **purity – degree to which a cluster contains words from one class only (best = 1)**
  - entropy – whether words from different classes are represented in the same cluster (best = 0)
  - 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
```

Intrinsic evaluation on word pairs: Analogy

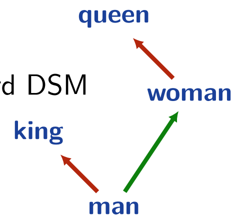
Mikolov *et al.* (2013b,a); Gladkova *et al.* (2016)

- 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)

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

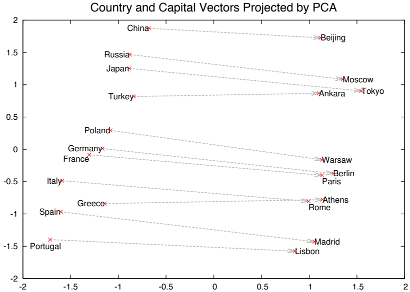
(Mikolov *et al.* 2013b, Tab. 1)

The 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.* 2014a)
FastText	82%	87%	(Mikolov *et al.* 2018)

Outline

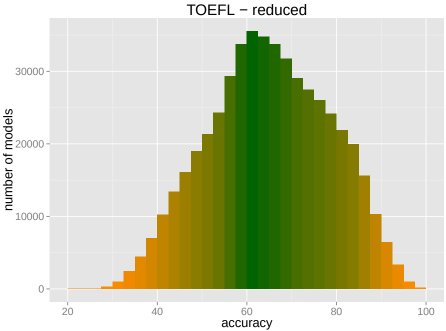
Making sense of evaluation results

Interpreting performance vs. picking the best run

1. **One model, many tasks** (Padó & Lapata 2007; Baroni & Lenci 2010; Pennington *et al.* 2014)
    - Novel DSM, one (or very few) settings tested on many tasks
    - Problem: not suitable for the exploration of a large parameter set, very limited coverage of interactions
2. **Incremental tuning** (Bullinaria & Levy 2007, 2012; Kiela & Clark 2014; Polajnar & Clark 2014)
    - Set parameter *a*, then *b*, then *c*
    - Problem: order dependent, very limited coverage of interactions
3. **Test all combinations** (Baroni *et al.* 2014a; Levy *et al.* 2015; Lapesa & Evert 2014)
    - Many tasks, many parameters, all combinations
    - Problem: many runs, interpreting results is a challenge

Lots of variation to make sense of...

TOEFL: 504k (!!!) runs (Lapesa & Evert 2014)



We need an interpretation methodology that:

- ... is able to identify robust trends, avoiding overfitting
- ... is able to capture parameter interactions

Outline

Linear regression to the rescue

- 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)

How to interpret the evaluation results?

Our proposal: linear regression

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, association score)

Motivation

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.

How to interpret the evaluation results?

Our proposal: linear regression

$$\boxed{\text{model performance} = \beta_0 + \beta_1 \cdot p_1 + \beta_2 \cdot p_2 + \beta_3 \cdot p_{1*2} + \ldots + \epsilon}$$

1. Adjusted R^2: proportion of variance explained by the model
   - ⇝ How well do we predict performance?
2. Feature ablation: proportion of variance explained by a parameter together with all its interactions
   - ⇝ Which parameters affect performance the most?
3. Model predictions: visualization of predicted performance
   - ⇝ What are the best parameter values?

How well do we predict performance?

A concrete example: TOEFL, SVD (504k data points)

accuracy ∼ ...

corpus	window	score	transformation	metric	n.dim	dim.skip	rel.index	accuracy
wacky	8	t-score	none	manhattan	700	0	dist	71.25
bnc	16	z-score	root	cosine	100	100	rank	75.00
wacky	16	MI	log	cosine	100	50	dist	77.50
bnc	8	frequency	none	cosine	900	50	rank	75.00
ukwac	16	MI	none	cosine	500	100	rank	81.25
bnc	8	tf.idf	root	cosine	300	100	rank	75.00
bnc	16	tf.idf	root	manhattan	300	100	dist	51.25
ukwac	2	tf.idf	log	manhattan	300	50	rank	53.75
ukwac	1	simple-ll	log	manhattan	500	100	dist	85.00

Model fit: Adj.R^2

Assumption: a good linear model acts as a “smoothing" algorithm which filters away random noise & captures robust trends.

How well do we predict performance?

A concrete example: TOEFL, SVD (504k data points)

accuracy ∼ corpus + window + score + transformation
+ metric + rel.index

corpus	window	score	transformation	metric	n.dim	dim.skip	rel.index	accuracy
wacky	8	t-score	none	manhattan	700	0	dist	71.25
bnc	16	z-score	root	cosine	100	100	rank	75.00
wacky	16	MI	log	cosine	100	50	dist	77.50
bnc	8	frequency	none	cosine	900	50	rank	75.00
ukwac	16	MI	none	cosine	500	100	rank	81.25
bnc	8	tf.idf	root	cosine	300	100	rank	75.00
bnc	16	tf.idf	root	manhattan	300	100	dist	51.25
ukwac	2	tf.idf	log	manhattan	300	50	rank	53.75
ukwac	1	simple-ll	log	manhattan	500	100	dist	85.00

Model fit:	**Adj.R^2**
basic	43%

Assumption: a good linear model acts as a "smoothing" algorithm which filters away random noise & captures robust trends.

How well do we predict performance?

A concrete example: TOEFL, SVD (504k data points)

```
accuracy ~ corpus + window + score + transformation
           + metric + rel.index + n.dim + dim.skip
```

corpus	window	score	transformation	metric	n.dim	dim.skip	rel.index	accuracy
wacky	8	t-score	none	manhattan	700	0	dist	71.25
bnc	16	z-score	root	cosine	100	100	rank	75.00
wacky	16	MI	log	cosine	100	50	dist	77.50
bnc	8	frequency	none	cosine	900	50	rank	75.00
ukwac	16	MI	none	cosine	500	100	rank	81.25
bnc	8	tf.idf	root	cosine	300	100	rank	75.00
bnc	16	tf.idf	root	manhattan	300	100	dist	51.25
ukwac	2	tf.idf	log	manhattan	300	50	rank	53.75
ukwac	1	simple-ll	log	manhattan	500	100	dist	85.00

Model fit:	Adj.R^2
basic	43%
& SVD	+24%

Assumption: a good linear model acts as a "smoothing" algorithm which filters away random noise & captures robust trends.

How well do we predict performance?

A concrete example: TOEFL, SVD (504k data points)

```
accuracy ~ corpus * window * score * transformation
           * metric * rel.index * n.dim * dim.skip
```

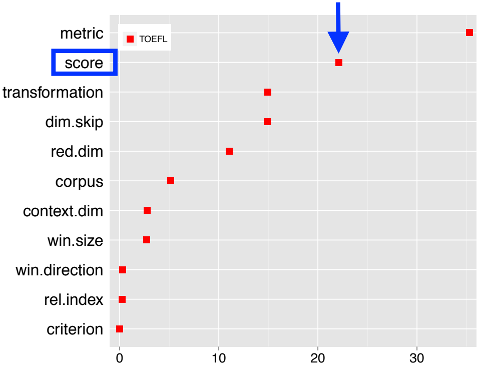
corpus	window	score	transformation	metric	n.dim	dim.skip	rel.index	accuracy
wacky	8	t-score	none	manhattan	700	0	dist	71.25
bnc	16	z-score	root	cosine	100	100	rank	75.00
wacky	16	MI	log	cosine	100	50	dist	77.50
bnc	8	frequency	none	cosine	900	50	rank	75.00
ukwac	16	MI	none	cosine	500	100	rank	81.25
bnc	8	tf.idf	root	cosine	300	100	rank	75.00
bnc	16	tf.idf	root	manhattan	300	100	dist	51.25
ukwac	2	tf.idf	log	manhattan	300	50	rank	53.75
ukwac	1	simple-ll	log	manhattan	500	100	dist	85.00

Model fit:	**Adj.R^2**
basic	43%
& SVD	+24%
& 2-way	+22%
Total:	**87%**

Assumption: a good linear model acts as a "smoothing" algorithm which filters away random noise & captures robust trends.

Which parameters affect performance the most?

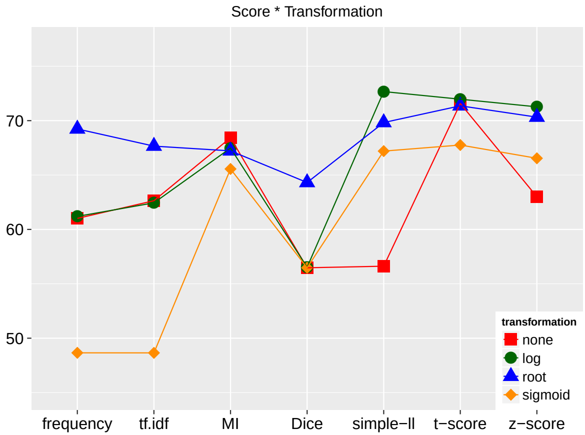
Feature ablation: parameters and interactions on TOEFL



Effect	R^2
score	10.53
score:transformation	7.42
score:metric	1.77
corpus:score	0.84
score:context.dim	0.64
other int. < 0.5	0.93
Feature ablation	**22.13**

Which parameters affect performance the most?

Interaction of score and transformation: effect plot



So, are there general trends? (Lapesa & Evert 2014)

Datasets: TOEFL, RG65, WordSim353, ESSLLI08 (and 3 other clust. datasets)

- Most explanatory parameters: similar across tasks/datasets
  - Simple-ll * Logarithmic Transformation, Cosine Distance
- Parameters that show variation: the amount and nature of shared context
  - Context window: 4 is a good compromise solution
  - SVD: always helps, and skipping the first dimensions (but not too many) generally helps
- Neighbor rank (almost) always better than distance
- Syntax (almost) never helps :((Lapesa & Evert 2017)

Contrasting semantic relations (Lapesa *et al.* 2014)

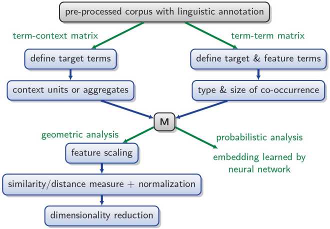
Datasets: Semantic Priming Project, GEK priming dataset

- **Semantic relations**
  - Paradigmatic (synonyms, antonyms, co-hyponyms) vs. Syntagmatic (phrasal associates, event associates)
- **Task: multiple choice**
- **Goal**: find the parameters which make the difference!
  - First SVD dimensions encode topical information, detrimental for paradigmatic relations (good to skip, also for TOEFL)
  - Syntagmatic relations: larger windows sizes. Co-occur, hence share context, but we need to enlarge the scope
- **Antonyms**: the least canonical paradigmatic
  - Larger windows, more relatedness like: antonyms co-occur (Justeson & Katz, 1992). Topic-shifting synonyms?
  - Less asymmetric (less difference between distance and rank)

Mid-lecture summary

- We introduced the coordinates of DSM evaluation
- We encountered (and started to get our hands dirty with) 3 standard tasks:
  - Multiple choice, prediction of similarity ratings, noun categorization
  - ☞ It is now your turn to practice, putting together all you learnt yesterday and the `wordspaceEval` datasets
- We also discussed the issue of DSM evaluation methodologies
  - Hopefully we persuaded you of how much variation parameter manipulation can introduce
  - ☞ maybe this motivates you even more to carry out a lot of experiments! So let us switch to RStudio now :)

Coming soon . . .



TOEFL dataset

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

Almuhareb Poesio

402 nouns, 21 classes
day ⟹ TIME
kiwi ⟹ FRUIT
kitten ⟹ ANIMAL
volleyball ⟹ GAME

BATTIG set

83 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

Rubenstein and Goodenough

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

WordSim

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

. . . but not yet, there is still something we need to talk about before turning to the practice session :)

DSM similarity & Linguistic Theory

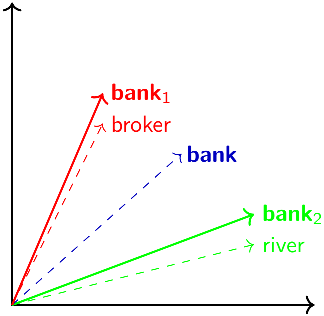
1. **Polysemy**
   - A textbook challenge, we will discuss the most intuitive solution
   - ☞ ... available in `wordspace`!
   - ☞ Code from the lecture and extensions in `hands_on_day4.R`
2. **Compositionality**
   - Above and below word level
   - ☞ Bonus evaluation dataset: derivational morphology in (Lazaridou *et al.* 2013)
   - ☞ Last part of `hands_on_day3.R`: perform your own standard tasks on `Lazaridou2013`
3. **Not all meaning is distributional**
   - Function words, proper names (literature pointers)

Great overview paper:
Distributional Semantics and Linguistic Theory (Boleda 2020)

Outline

Polysemy in DSMs

- Problem: DSM vectors conflate contexts from different senses of a word
  - contexts of "bank": money, river, account, swim, ...
  - vectors are displaced suboptimally (far from everything)



Polysemy in DSMs

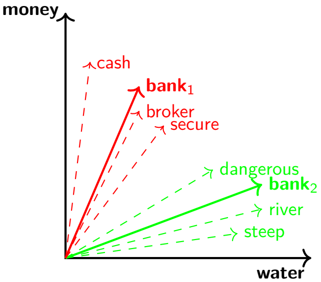
Observation: DSM vectors conflate contexts from word senses

- Solution: build a representation for each instance of the word we want to disambiguate (Schütze 1998)

sentence vectors

Target: bank

bank$_1$: *The broker went to the bank to secure his cash*
bank$_2$: *The river bank was steep and dangerous*



Application: word sense disambiguation
... can you think about another situation in which we may need it?

Context vectors: can we do it in wordspace?

Yes :D

```
library(wordspace)
# S1: ‘‘Cats and dogs need their time’’
s1 <- "cat and dog need their time"
# S2: ‘‘Time is the cause not the effect’’
s2 <- "time is the cause not the effect"
# Ingredients: vectors for individual words
>TT <- DSM_TermTermMatrix
>TT
       breed tail feed kill important explain likely
cat       84   17    8   38         0       2      0
dog      579   14   32   63         1       2      2
animal    45   11   86  136        13       5      4
time      19    8   29  134        94      44    100
reason     1    0    1   18        71     140     39
cause      0    1    0    3        55      35     51
effect     0    1    1    6        62      37     14
```

Context vectors: can we do it in `wordspace`?

Yes :D

"cats and dogs need their time"

```
> context.vectors(TT, s1)
     breed tail feed     kill important explain likely
1 227.3333   13   23 78.33333  31.66667      16     34
# context.vectors() is taking the average of the values in each cell
> (TT['cat','breed']+TT['dog','breed']+TT['time','breed'])/3
227.3333
```

"time is the cause not the effect"

```
round(context.vectors(TT, s2),3)
  breed  tail feed   kill important explain likely
1 6.333 3.333   10 47.667    70.333  38.667     55
```

Context vectors: can we do it in wordspace?

Almost there...

```
# context.vectors() can also take a list as an input
contexts <- round(context.vectors(TT, c(s1, s2)),2)
# The output is a matrix, let's give it better rownames first
rownames(contexts) <- c("s1", "s2")
# ...and then append it to our original matrix
TT <- rbind(TT, contexts)
TT
        breed  tail feed   kill important explain likely
cat     84.00 17.00    8  38.00      0.00    2.00      0
dog    579.00 14.00   32  63.00      1.00    2.00      2
animal  45.00 11.00   86 136.00     13.00    5.00      4
time    19.00  8.00   29 134.00     94.00   44.00    100
reason   1.00  0.00    1  18.00     71.00  140.00     39
cause    0.00  1.00    0   3.00     55.00   35.00     51
effect   0.00  1.00    1   6.00     62.00   37.00     14
s1     227.33 13.00   23  78.33     31.67   16.00     34
s2       6.33  3.33   10  47.67     70.33   38.67     55
```

Context vectors: can we do it in wordspace?

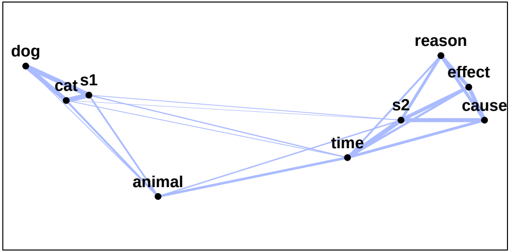
And what now?

```
# We can do all the cool things we are used to do with DSM matrices
# Nearest neighbors...
nearest.neighbours(TT, c("s1", "s2"), n=6)
$s1
     cat      dog   animal     time       s2    cause
14.31016 17.16200 55.27587 62.66470 67.81707 77.90557

$s2
    time    cause   effect   reason   animal       s1
18.85097 25.19348 31.51682 40.83768 60.61621 67.81707
```

Context vectors: can we do it in `wordspace`?

```
# And a semantic map!
plot(dist.matrix(TT))
```



`hands_on_day_4.R` also contains an example for the *bank* polysemy, with word2vec vectors. If you fell in love with centroids the bonus exercise in `schuetze1998.R` (word sense disambiguation, advanced) is perfect for you!

Polysemy in DSMs: contextualized word embeddings

A little detour in embeddingland: BERT

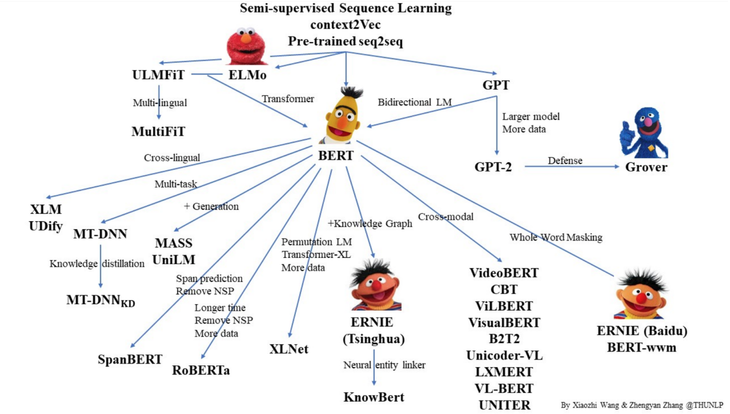
Next step: one contextualized representation per token

The$_1$, broker$_1$, went$_1$, to$_2$, the$_1$, bank$_1$, I$_2$, swam$_2$, to$_2$, the$_2$, bank$_2$, The$_3$, river$_3$, bank$_3$, is$_3$, steep$_3$

- Bidirectional Encoder Representations from Transformers
- **Most popular embeddings right now. Why?**
  - Multilingual and easily fine-tuned for specific tasks (e.g., question answering, sentiment analysis)
  - Google open-source NLP framework (2018) (`https://github.com/google-research/bert`)
    - Pre-trained on Wikipedia (2.5B tokens) + Google Books (800M tokens)

Polysemy in DSMs: contextualized word embeddings

BERT & other Animals



Problem: some tasks (e.g., those from) require lemma-level representations, which need to be reconstructed "backwards"

Outline

Compositionality

Can we capture it in DS?

- Formally: compositionality implies some operator $\bigoplus$ such that

$$\text{meaning}(w_1 w_2) = \text{meaning}(w_1) \bigoplus \text{meaning}(w_2)$$

- CDSM recipe
  - Distributional vectors for meaning(w_1) and meaning(w_2)
  - Operators: mathematical stategies to combine w_1 and w_2 to *predict* a vector representation for $w_1 w_2$
    - ⋆ vector addition
    - ⋆ vector multiplication
    - ⋆ nonlinear operations learned by neural networks
- Problem: some words (e.g., not) are themselves more like operators than points in space

Great overview paper: Frege in space: a program for compositional distributional semantics (Baroni *et al.* 2014b)

Compositionality with distributional vectors

Additive and Multiplicative Models (Mitchell and Lapata, 2010)

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0
problem	2	15	7	9	1

$$p = u + v$$

predicted(practical difficulty) = **practical** + **difficulty** = [1 14 6 14 4]

$$p = u \odot v$$

predicted(practical difficulty) = **practical** $\odot$ **difficulty** = [0 48 8 40 0]

What is your intuition about the effect of multiplication? Have you already seen it as an ingredient of something else?

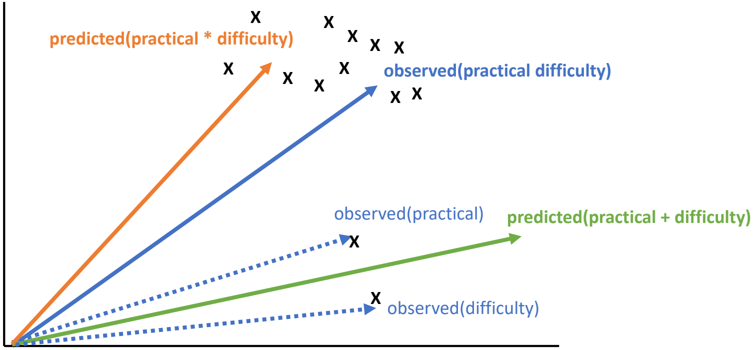
How do I know my composed representations are "good"?

Evaluation, again :)

1. **Qualitative inspection of nearest neighbors**
   - Which neighbors "make more sense" ?
     - practical $+$ difficulty or practical $\odot$ difficulty ?
2. **Quantitative evaluation**
   - Collect a vector for "practical difficulty" in (obviously the same) corpus: **observed(practical difficulty)**
   - observed(practical difficulty) $\approx$ predicted(practical difficulty)
     - Which of the two produces a better approximation?
     - practical $+$ difficulty or practical $\odot$ difficulty
   - Evaluation metric
     - distance(predicted,observed) (Lazaridou *et al.* 2013)
     - rank(predicted,observed) (Baroni & Zamparelli 2010; Padó *et al.* 2016)

How do I know my composed representations are “good”?

Observed vs. Predicted vector



rank(predicted(practical + difficulty)) = 5 < rank(predicted(practical * difficulty)) = 10

distance(predicted(practical * difficulty)) < distance(predicted(practical + difficulty))

Adjective-noun composition (Baroni & Zamparelli 2010)

Starting point: observed AN vectors

- **Input**: triples of {observed(AN), A, N}
  - {bad luck, bad, luck}, {red cover, red, cover}, etc.
  - 36 adjectives (size, color, temporal, etc.)

bad luck	*electronic communities*	*historical map*
bad	electronic storage	topographical
bad weekend	electronic transmission	atlas
good spirit	purpose	historical material
important route	*nice girl*	*little war*
important transport	good girl	great war
important road	big girl	major war
major road	guy	small war
red cover	*special collection*	*young husband*
black cover	general collection	small son
hardback	small collection	small daughter
red label	archives	mistress

- **Methods**: increasing computational complexity
  - No learning (additive, multiplicative)
  - ☞ heavy learning: learns matrix A by comparing AN and N

Adjective-noun composition in Baroni & Zamparelli (2010)

Best method: adjectives as matrices. Observed(AN) vs. predicted(AN): neighbors

SIMILAR			DISSIMILAR		
adj N	*obs. neighbor*	*pred. neighbor*	*adj N*	*obs. neighbor*	*pred. neighbor*
common understanding	common approach	common vision	American affair	Am. development	Am. policy
different authority	diff. objective	diff. description	current dimension	left (a)	current element
different partner	diff. organisation	diff. department	good complaint	current complaint	good beginning
general question	general issue	*same*	great field	excellent field	gr. distribution
historical introduction	hist. background	*same*	historical thing	different today	hist. reality
necessary qualification	nec. experience	*same*	important summer	summer	big holiday
new actor	new cast	*same*	large pass	historical region	large dimension
recent request	recent enquiry	*same*	special something	little animal	special thing
small drop	droplet	drop	white profile	chrome (n)	white show
young engineer	young designer	y. engineering	young photo	important song	young image

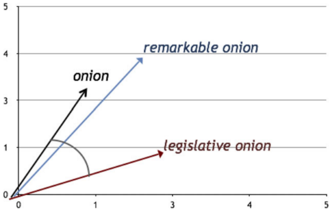
Table 4: Left: nearest neighbors of observed and *alm*-predicted ANs (excluding each other) for a random set of ANs where rank of observed w.r.t. predicted is 1. Right: nearest neighbors of predicted and observed ANs for random set where rank of observed w.r.t. predicted is $\geq 1K$.

How about unattested AN combinations?

Capturing Semantically Deviant AN Combinations (Vecchi *et al.* 2017)

Can we use compositional DSMs to tell, among equally unattested AN, which one is semantically less plausible?

The *composed vectors* for semantically deviant (human rated) combinations will be farther away from the head noun than the acceptable ones



... they test other measures (e.g., neighbors density, vector length) as well as different composition methods: have a look at the paper!

How about unattested AN combinations?

Capturing Semantically Deviant AN Combinations (Vecchi *et al.* 2017)

Can we use compositional DSMs to tell, among equally unattested AN, which one is semantically less plausible?

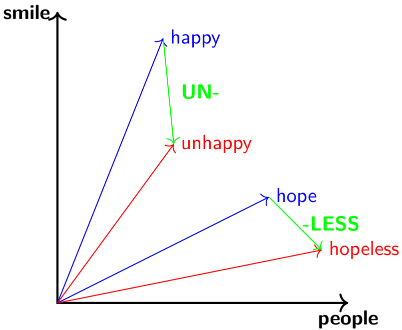
Qualitative inspection: the *composed vectors* of semantically acceptable pairs have plausible nearest neighbors

a.	*angry lamp	{ *shocked*, *fearful*, *angry*, *defiant* }
b.	*nuclear fox	{ *nuclear*, *nuclear arm*, *nuclear development*, *nuclear expert* }
c.	warm garlic	{ *green salad*, *wild mushroom*, *sauce*, *green sauce* }
d.	spectacular striker	{ *goal*, *crucial goal*, *famous goal*, *amazing goal* }

`hands_on_day_4.R` (part 2) contains an implementation of vector addition and multiplication in `wordspace`. Have fun chasing the strangest AN combinations! And other combinations, as well

Compositionality below word level

Can we use compositional DSMs to investigate the meaning of derivational patterns?



- Starting point: vectors for base and derived words.
- Two strategies:
  - ☞ learn the **semantic shifts** with compositional methods
  - investigate **properties** of the patterns → semantic relations
    - ★ zero-nominalizations as hyponyms of the base verb (Varvara *et al.* 2021)
    - ★ un- as antonyms of the base nouns

The DS of Derivational Morphology (Lazaridou *et al.* 2013)

1. **Input**: derived/stem vector pairs for each affix
   - un-: unfaithful/faithful, unbiased/biased, unwell/well
   - -ly: true/truly, mad/madly, deep/deeply
2. **Goal: build one representation per affix**
   - No (well, little) learning (additive and multiplicative)
     - un- = centroid(unfaithful, unbiased, unwell, etc.)
   - Increasingly complex learning
     - Parameters set during training to optimize composition, affixes as matrices (cf. adjectives)
3. **Prediction & Evaluation**
   - Apply affix to unseen base: predicted(derived) vs. observed(derived). Who did it best?
     - Simplest (additive) & most complex (lexical functional, theoretically motivated): comparable
     - Cf. Padó *et al.* (2016) for German: simplest composition methods work better!

The DS of Derivational Morphology (Lazaridou *et al.* 2013)

Dataset

Affix	Stem/Der. POS	Training Items	HQ/Tot. Test Items	Avg. SDR
-able	verb/adj	177	30/50	5.96
-al	noun/adj	245	41/50	5.88
-er	verb/noun	824	33/50	5.51
-ful	noun/adj	53	42/50	6.11
-ic	noun/adj	280	43/50	5.99
-ion	verb/noun	637	38/50	6.22
-ist	noun/noun	244	38/50	6.16
-ity	adj/noun	372	33/50	6.19
-ize	noun/verb	105	40/50	5.96
-less	noun/adj	122	35/50	3.72
-ly	adj/adv	1847	20/50	6.33
-ment	verb/noun	165	38/50	6.06
-ness	adj/noun	602	33/50	6.29
-ous	noun/adj	157	35/50	5.94
-y	noun/adj	404	27/50	5.25
in-	adj/adj	101	34/50	3.39
re-	verb/verb	86	27/50	5.28
un-	adj/adj	128	36/50	3.23
tot	*/*	6549	623/900	5.52

7000 base/derived pairs from CELEX, 18 patterns, training vs. test (further annotated for base/derived relatedness and vector quality)

Outline

Not all Semantic Knowledge is Distributional

Proper names "answer the purpose of showing what thing it is that we are talking about but not of telling anything about it" (Mill, 1843)

- Intuition: instances of categories such as PER, ORG, etc.
- Herbelot (2015), standard DSMs: category → instance
  - "… upon encountering the name *Mr Darcy* for the first time in the novel, a reader will attribute it the representation of the concept `man` and subsequently specialise it as per the linguistic contexts in which the name appears"
- Westera *et al.* (2021), embeddings: instance → category

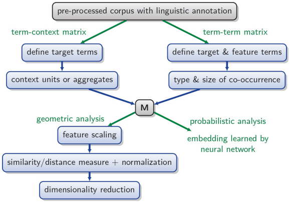
Function words: some pointers

- Baroni *et al.* (2012) on quantifiers/entailment, Bernardi *et al.* (2013) on determiners, Hole & Padó (2021) on the polysemy of the German reflexive *sich*

Wrapping up

- Distributional semantics allows us to represent (and compare) a quite heterogeneous selection of "linguistic objects":
  - Subword units (e.g., derivational affixes)
  - Words (content words, proper names, function words)
  - Phrases (e.g., AN)
  - Entire sentences
- This is fascinating and promising, but also challenging
  - On top of the DSM parameters, also other experimental choices (e.g., composition. methods)
- . . . and this is exactly the fun of distributional semantics (at least for us :))
  - ☞ Now it is finally your turn to have fun

It is practice session time!



TOEFL dataset

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

Almuhareb Poesio

402 nouns, 21 classes
day ⟹ TIME
kiwi ⟹ FRUIT
kitten ⟹ ANIMAL
volleyball ⟹ GAME

BATTIG set

83 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

Rubenstein and Goodenough

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

WordSim

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

Affix	Stem/Der. POS	Training Items	HQ/Tot. Test Items	Avg. SDR
-able	verb/adj	177	30/50	5.96
-al	noun/adj	245	41/50	5.88
-er	verb/noun	824	33/50	5.51
-ful	noun/adj	53	42/50	6.11
-ic	noun/adj	280	43/50	5.99
-ion	verb/noun	637	38/50	6.22
-ist	noun/noun	244	38/50	6.16
-ity	adj/noun	372	33/50	6.19
-ize	noun/verb	105	40/50	5.96

References I

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

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 and Zamparelli, Roberto (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Baroni, Marco; Bernardi, Raffaella; Do, Ngoc-Quynh; Shan, Chung-chieh (2012). Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32, Avignon, France. Association for Computational Linguistics.

References II

Baroni, Marco; Dinu, Georgiana; Kruszewski, Germán (2014a). 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.

Baroni, Marco; Bernardi, Raffaelle; Zamparelli, Roberto (2014b). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology (LiLT)*, **9**(6), 5–109.

Bernardi, Raffaella; Dinu, Georgiana; Marelli, Marco; Baroni, Marco (2013). A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 53–57, Sofia, Bulgaria. Association for Computational Linguistics.

Boleda, Gemma (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, **6**(1), 213–234.

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 III

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.

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.

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*.

Herbelot, Aurélie (2015). Mr darcy and mr toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 151–161, London, UK. Association for Computational Linguistics.

References IV

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.

Hole, Daniel and Padó, Sebastian (2021). Distributional analysis of function words. To appear in Proceedings of the 13th International Tbilisi Symposium on Language, Logic and Computation.

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.

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.

References V

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.

Lapesa, Gabriella and Evert, Stefan (2017). Large-scale evaluation of dependency-based DSMs: Are they worth the effort? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 394–400, Valencia, Spain. Association for Computational Linguistics.

Lapesa, Gabriella; Evert, Stefan; Schulte im Walde, Sabine (2014). Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 160–170, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

References VI

Lazaridou, Angeliki; Marelli, Marco; Zamparelli, Roberto; Baroni, Marco (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1526, Sofia, Bulgaria. Association for Computational Linguistics.

Levy, Omer; Goldberg, Yoav; Dagan, Ido (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, **3**, 211–225.

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*.

References VII

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.

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

Padó, Sebastian; Herbelot, Aurélie; Kisselew, Max; Šnajder, Jan (2016). Predictability of distributional semantics in derivational word formation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1285–1296, Osaka, Japan. The COLING 2016 Organizing Committee.

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

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.

References VIII

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.

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

Sahlgren, Magnus (2008). The distributional hypothesis. *Italian Journal of Linguistics*, **20**.

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.

Schütze, Hinrich (1998). Automatic word sense discrimination. *Computational Linguistics*, **24**(1), 97–123.

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

Varvara, Rossella; Lapesa, Gabriella; Padó, Sebastian (2021). Grounding semantic transparency in context: A distributional semantic study on German event nominalizations. *Morphology*.

References IX

Vecchi, Eva M.; Marelli, Marco; Zamparelli, Roberto; Baroni, Marco (2017). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, **41**(1), 102–136.

Westera, Matthijs; Gupta, Abhijeet; Boleda, Gemma; Padó, Sebastian (2021). Distributional models of category concepts based on names of category members. *Cognitive Science*. Accepted for publication. Preprint available at https://nlpado.de/ sebastian/pub/papers/WesteraEtal2021.pdf.