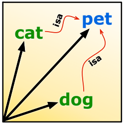
Distributional Semantic Models

Part 3: Evaluation – is my DSM "good"?

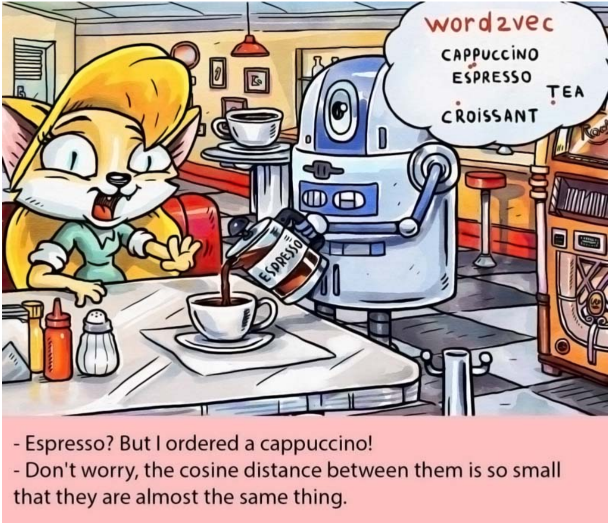
Stephanie Evert[1] & Gabriella Lapesa[4]
with Alessandro Lenci[2] and Marco Baroni[3]

[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


word2vec
CAPPUCCINO
ESPRESSO
TEA
CROISSANT
ESPRESSO
- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

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)?

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!

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

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 & 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?

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

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*)

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*)

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)

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)

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)

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)

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*

```
> library(wordspaceEval)
> head(TOEFL80)
```

The TOEFL synonym task

- The TOEFL dataset (80 items)
  - Target: *show*
    Candidates: *demonstrate*, *publish*, *repeat*, *postpone*

```
> library(wordspaceEval)
> head(TOEFL80)
```

The TOEFL synonym task

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

```
> library(wordspaceEval)
> head(TOEFL80)
```

The TOEFL synonym task

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

```
> library(wordspaceEval)
> head(TOEFL80)
```

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%

And you?

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

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%

And you?

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

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 (`https://aclweb.org/aclwiki/Similarity_(State_of_the_art)`)
  - Term-Document: Classic LSA (Landauer & Dumais 1997): 64.4%
  - Dependency-filtered Padó and Lapata's (2007): 73.0%
  - Depedency-typed (Baroni & Lenci 2010): 76.9%
  - Term-term Bullinaria & Levy (2012) , 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

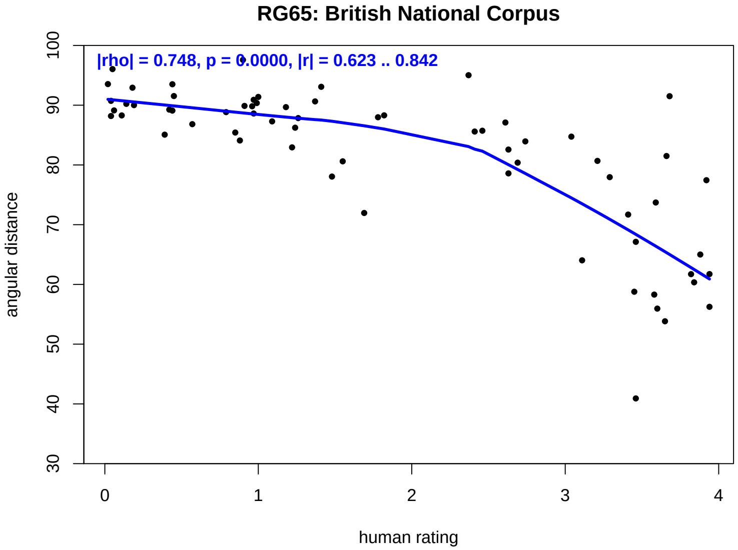
weapon – *secret*: 6.06

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 judgments

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

Semantic similarity judgments: example



Semantic similarity judgments: results

Results on RG65 task (Pearson):

- Dependency-filtered, BNC: Padó and Lapata (2007): 0.62
- Dependency-filtered, Web data (Herdağdelen *et al.* 2009)
  - without SVD reduction: 0.69
  - with SVD reduction: 0.80
- Dependency typed (Baroni & Lenci 2010): 0.82
- Term-term + some magic (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

ESSLI08 (on focus today)

44 nouns, 6 classes

potato $\Longrightarrow$ GREEN

hammer $\Longrightarrow$ TOOL

car $\Longrightarrow$ VEHICLE

peacock $\Longrightarrow$ BIRD

BATTIG set

82 nouns, 10 classes

chicken $\Longrightarrow$ BIRD

bear $\Longrightarrow$ LAND_MAMMAL

pot $\Longrightarrow$ KITCHENWARE

oak $\Longrightarrow$ TREE

Almuhareb & Poesio

402 nouns, 21 classes

day $\Longrightarrow$ TIME

kiwi $\Longrightarrow$ FRUIT

kitten $\Longrightarrow$ ANIMAL

volleyball $\Longrightarrow$ GAME

MITCHELL set

60 nouns, 12 classes

ant $\Longrightarrow$ INSECT

carrot $\Longrightarrow$ VEGETABLE

train $\Longrightarrow$ VEHICLE

cat $\Longrightarrow$ 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*)

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

Noun categorization: the ESSLLI 2008 dataset

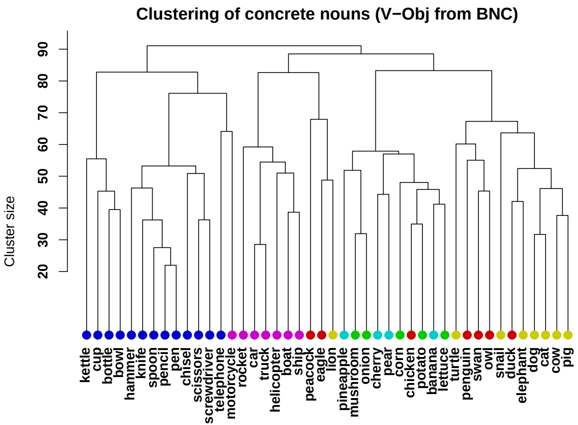
Dataset of 44 concrete nouns (ESSLLI 2008 Shared Task)

- 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*)

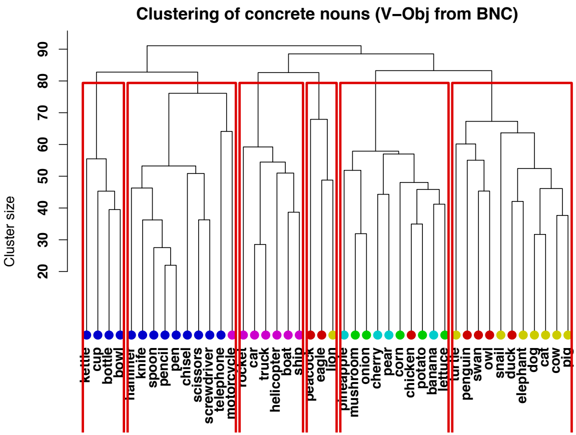
- DSMs operationalizes categorization as a clustering task
  1. for each noun w_i in the dataset, take its vector $\mathbf{w}_i$
  2. use a clustering method to group similar vectors $\mathbf{w}_i$
  3. evaluate whether clusters correspond to gold-standard semantic classes (purity, entropy, ...)

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

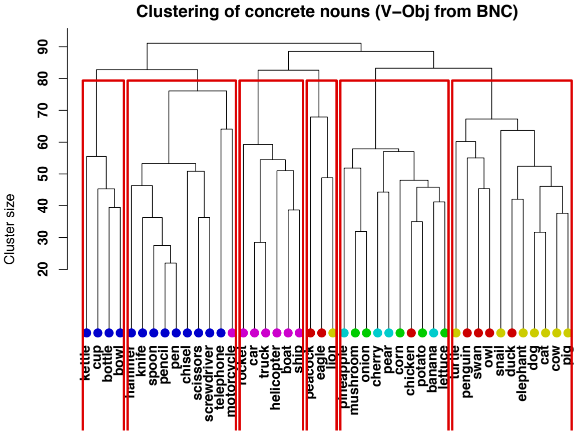
Noun categorization: example



Noun categorization: example

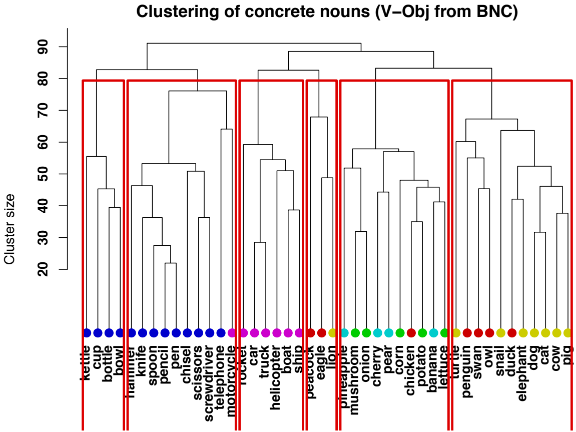


Noun categorization: example



- majority labels: tools, tools, vehicles, birds, greens, animals
- correct: 4/4, 9/10, 6/6, 2/3, 5/10, 7/11

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

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*	
Pattern-based (Katrenko)	89	13	100	0	80	59	197
Term-term (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

`Katrenko`, `Peirsman`: 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* : **???**

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* : **???**

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* : **???**

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* : **???**

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*

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

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

Intrinsic evaluation on word pairs: Analogy

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

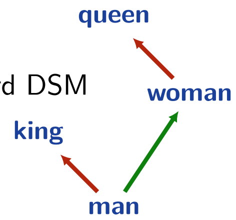
- Task: solve analogy problems such as
  - *man* : *woman* :: *king* : *queen*
  - *France* : *Paris* :: *Bulgaria* : *Sofia*
  - *learn* : *learned* :: *go* : *went*
  - *dog* : *animal* :: *strawberry* : *fruit*

- Approach 1: build DSM on word pairs as targets

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

- 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.*

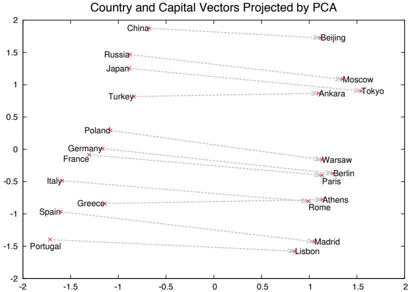
Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	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

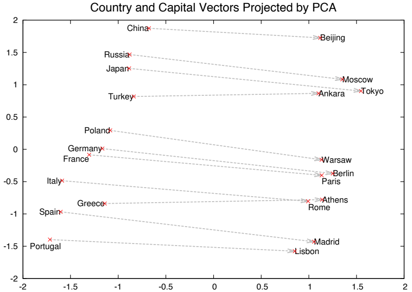


(Mikolov *et al.* 2013a, Fig. 2)

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)

Outline

Making sense of evaluation results

Interpreting performance vs. picking the best run

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

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

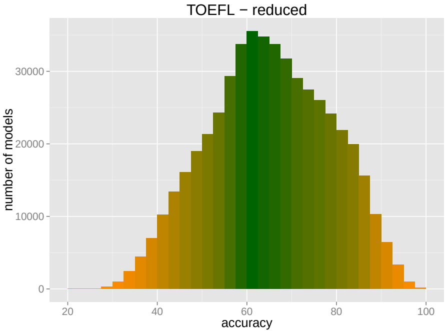
Making sense of evaluation results

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.* 2014; 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

DSM evaluation: coordinates

- Tasks & Datasets

DSM evaluation in theory and with `wordspaceEval`

- Multiple choice
- Prediction of similarity ratings
- Noun categorization

Methodology for DSM Evaluation

- Previous work
- Interpreting DSM performance with linear regression

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

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

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
   $\rightsquigarrow$ How well do we predict 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?

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
   - $\leadsto$ How well do we predict performance?
2. Feature ablation: proportion of variance explained by a parameter together with all its interactions
   - $\leadsto$ Which parameters affect performance the most?
3. Model predictions: visualization of predicted performance
   - $\leadsto$ 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
```

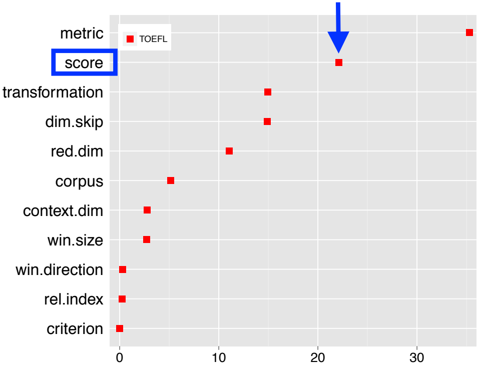
corpus	window	score	transformation	metric	n.dim	dim.skip	rel.index	accuracy
wacky	8	t-score	none	manhattan	700	0	dist	71.25
bnc	16	z-score	root	cosine	100	100	rank	75.00
wacky	16	MI	log	cosine	100	50	dist	77.50
bnc	8	frequency	none	cosine	900	50	rank	75.00
ukwac	16	MI	none	cosine	500	100	rank	81.25
bnc	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?

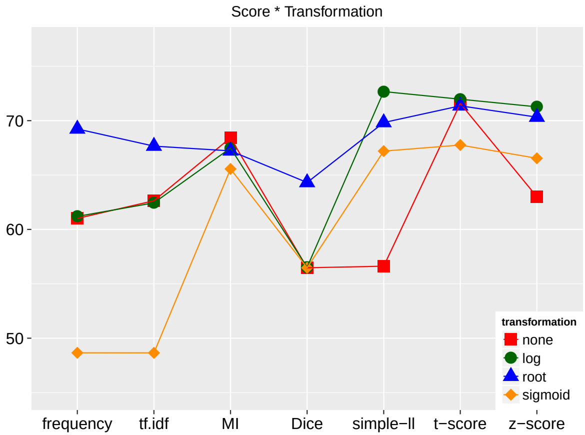
Feature ablation: parameters and interactions on TOEFL



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)

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

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

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

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

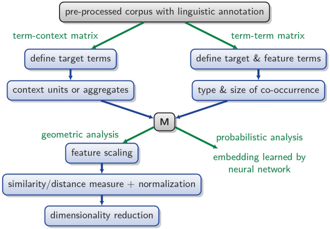
Datasets: TOEFL, RG65, WordSim353, ESSLLI08 (and 3 other clust. datasets)

- 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)

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 :)

Your turn now!



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

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; 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.

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.

References II

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*.

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.

References III

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

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.

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.

References IV

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.

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 V

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.

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.

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**.

References VI

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.