Distributional Semantic Models

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

Stefan Evert\textsuperscript{1} & Gabriella Lapesa\textsuperscript{4}  
with Alessandro Lenci\textsuperscript{2} and Marco Baroni\textsuperscript{3}

\textsuperscript{1}Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany  
\textsuperscript{2}University of Pisa, Italy  
\textsuperscript{3}University of Trento, Italy  
\textsuperscript{4}University of Stuttgart, Germany

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

Copyright © 2009–2021 Evert, Lapesa, Lenci & Baroni | Licensed under CC-by-sa version 3.0
“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

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

DS beyond NLP: Linguistic evaluation
Polysemy
Compositionality
Non distributional meaning
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 similarity/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 & 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 similarity/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)

© Evert/Lenci/Baroni/Lapesa (CC-by-sa)
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 similarity/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 similarity/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

Intrinsic 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)
  - targets = lemma + POS code (e.g. `white_J`)
  - compatible with evaluation tasks included in package

```r
library(wordspace)

M <- DSM_Vectors
nearest.neighbours(M, "walk_V")

  amble_V  19.4
  stroll_V  21.8
  traipse_V  21.8
  potter_V  22.6
  tramp_V   22.9
  saunter_V  23.5
  wander_V   23.7
  trudge_V   23.8
  leisurely_R 26.2
  saunter_N  26.4

# you can also try `white`, `apple` and `kindness`
```
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

DS beyond NLP: Linguistic evaluation
Polysemy
Compositionality
Non distributional meaning
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

```r
> 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*

```r
> 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

```r
> 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 (t) and of the candidates (c₁...cₙ)
  2. measure the distance between t and cᵢ, with 1 ≤ i ≤ n
  3. select cᵢ with the shortest distance in space from t

```r
> 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
  - 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

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

DS beyond NLP: Linguistic evaluation
Polysemy
Compositionality
Non distributional meaning
Semantic similarity judgments

<table>
<thead>
<tr>
<th>RG65</th>
<th>WordSim353</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>65 pairs, rated from 0 to 4</strong></td>
<td><strong>353 pairs, rated from 1 to 10</strong></td>
</tr>
<tr>
<td><em>gem</em> – <em>jewel</em>: 3.94</td>
<td><em>announcement</em> – <em>news</em>: 7.56</td>
</tr>
<tr>
<td><em>grin</em> – <em>smile</em>: 3.46</td>
<td><em>weapon</em> – <em>secret</em>: 6.06</td>
</tr>
<tr>
<td><em>fruit</em> – <em>furnace</em>: 0.05</td>
<td><em>travel</em> – <em>activity</em>: 5.00</td>
</tr>
</tbody>
</table>

▶ DSMs vs. Ratings: operationalization

1. for each test pair \((w_1, w_2)\), take vectors \(w_1\) and \(w_2\)
2. measure the distance (e.g. cosine) between \(w_1\) and \(w_2\)
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

RG65: British National Corpus

\[ |\rho| = 0.748, \ p = 0.0000, \ |r| = 0.623 \ldots 0.842 \]
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?

```r
> 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(RG65, M, convert=FALSE, details=TRUE))
```

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

**DS beyond NLP: Linguistic evaluation**
  Polysemy
  Compositionality
  Non distributional meaning
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**
    ★ $\lambda y : \text{FOOD} \ \lambda x : \text{ANIMATE} \ [\text{eat}(x, y)]$
# Noun categorization: datasets

<table>
<thead>
<tr>
<th>DSM evaluation in theory and with wordspaceEval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun categorization</td>
</tr>
</tbody>
</table>

## ESSLLI08 (on focus today)

<table>
<thead>
<tr>
<th>44 nouns, 6 classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>potato $\rightarrow$ GREEN</td>
</tr>
<tr>
<td>hammer $\rightarrow$ TOOL</td>
</tr>
<tr>
<td>car $\rightarrow$ VEHICLE</td>
</tr>
<tr>
<td>peacock $\rightarrow$ BIRD</td>
</tr>
</tbody>
</table>

## BATTIG set

<table>
<thead>
<tr>
<th>82 nouns, 10 classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>chicken $\rightarrow$ BIRD</td>
</tr>
<tr>
<td>bear $\rightarrow$ LAND_MAMMAL</td>
</tr>
<tr>
<td>pot $\rightarrow$ KITCHENWARE</td>
</tr>
<tr>
<td>oak $\rightarrow$ TREE</td>
</tr>
</tbody>
</table>

## Almuhareb & Poesio

<table>
<thead>
<tr>
<th>402 nouns, 21 classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>day $\rightarrow$ TIME</td>
</tr>
<tr>
<td>kiwi $\rightarrow$ FRUIT</td>
</tr>
<tr>
<td>kitten $\rightarrow$ ANIMAL</td>
</tr>
<tr>
<td>volleyball $\rightarrow$ GAME</td>
</tr>
</tbody>
</table>

## MITCHELL set

<table>
<thead>
<tr>
<th>60 nouns, 12 classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ant $\rightarrow$ INSECT</td>
</tr>
<tr>
<td>carrot $\rightarrow$ VEGETABLE</td>
</tr>
<tr>
<td>train $\rightarrow$ VEHICLE</td>
</tr>
<tr>
<td>cat $\rightarrow$ ANIMAL</td>
</tr>
</tbody>
</table>
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)

```r
> 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 $w_i$
  2. use a clustering method to group similar vectors $w_i$
  3. evaluate whether clusters correspond to gold-standard semantic classes (purity, entropy, ...)

```
> ESSLLI08_Nouns[seq(1,40,5), ]
```
Noun categorization: example
Noun categorization: example

Clustering of concrete nouns (V−Obj from BNC)

Cluster size

- 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%
Noun categorization: example

Clustering of concrete nouns (V−Obj from BNC)

- 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

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

\[ 3 \sum_{i=1}^{3} \text{Purity}_i - 3 \sum_{i=1}^{3} \text{Entropy}_i \]
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**

<table>
<thead>
<tr>
<th>model</th>
<th>6-way</th>
<th>3-way</th>
<th>2-way</th>
<th>global</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>E</td>
<td>P</td>
<td>E</td>
</tr>
<tr>
<td>Katrenko</td>
<td>89</td>
<td>13</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Peirsman+</td>
<td>82</td>
<td>23</td>
<td>84</td>
<td>34</td>
</tr>
<tr>
<td>dep-typed (DM)</td>
<td>77</td>
<td>24</td>
<td>79</td>
<td>38</td>
</tr>
<tr>
<td>dep-filtered (DM)</td>
<td>80</td>
<td>28</td>
<td>75</td>
<td>51</td>
</tr>
<tr>
<td>window (DM)</td>
<td>75</td>
<td>27</td>
<td>68</td>
<td>51</td>
</tr>
<tr>
<td>Peirsman−</td>
<td>73</td>
<td>28</td>
<td>71</td>
<td>54</td>
</tr>
<tr>
<td>Shaoul</td>
<td>41</td>
<td>77</td>
<td>52</td>
<td>84</td>
</tr>
</tbody>
</table>

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

**And you?**

```r
> 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
▶ \textit{man} : \textit{woman} :: \textit{king} : ???

© Evert/Lenci/Baroni/Lapesa (CC-by-sa)
Intrinsic evaluation on word pairs: Analogy

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

- Task: solve analogy problems such as
  - \( \text{man} : \text{woman} :: \text{king} : \text{queen} \)
  - \( \text{France} : \text{Paris} :: \text{Bulgaria} : \text{???} \)
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 : ???

© Evert/Lenci/Baroni/Lapesa (CC-by-sa) DSM Tutorial – Part 3  wordspace.collocations.de 24 / 80
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
  ▶ \( \text{man} : \text{woman} \quad \text{::} \quad \text{king} : \text{queen} \)
  ▶ \( \text{France} : \text{Paris} \quad \text{::} \quad \text{Bulgaria} : \text{Sofia} \)
  ▶ \( \text{learn} : \text{learned} \quad \text{::} \quad \text{go} : \text{went} \)
  ▶ \( \text{dog} : \text{animal} \quad \text{::} \quad \text{strawberry} : \text{fruit} \)
Intrinsic evaluation on word pairs: Analogy
Mikolov et al. (2013b,a); Gladkova et al. (2016)

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

- Approach 1: build DSM on word pairs as targets
  $$\min_x d(\mathbf{v}_{\text{man:woman}}, \mathbf{v}_{\text{king}:x})$$
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
  \[
  \min_x d(v_{\text{man}:\text{woman}}, v_{\text{king}:x})
  \]

- Approach 2: use vector operations in single-word DSM
  \[
  v_{\text{queen}} \approx v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}
  \]
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.

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

(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

![Country and Capital Vectors Projected by PCA](image)

(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

<table>
<thead>
<tr>
<th>model</th>
<th>syntactic</th>
<th>semantic</th>
</tr>
</thead>
<tbody>
<tr>
<td>word2vec</td>
<td>64%</td>
<td>55%</td>
</tr>
<tr>
<td>DSM</td>
<td>43%</td>
<td>60%</td>
</tr>
<tr>
<td>FastText</td>
<td>82%</td>
<td>87%</td>
</tr>
</tbody>
</table>
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

DS beyond NLP: Linguistic evaluation
  Polysemy
  Compositionality
  Non distributional meaning
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.* 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

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

DS beyond NLP: Linguistic evaluation
  Polysemy
  Compositionality
  Non distributional meaning
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

\[
\text{model performance} = \beta_0 + \beta_1 \cdot p_1 + \beta_2 \cdot p_2 + \beta_3 \cdot p_{1*2} + \ldots + \epsilon
\]
How to interpret the evaluation results?

Our proposal: linear regression

\[
\text{model performance} = \beta_0 + \beta_1 \cdot p_1 + \beta_2 \cdot p_2 + \beta_3 \cdot p_1 \cdot p_2 + \ldots + \epsilon
\]

1. Adjusted $R^2$: proportion of variance explained by the model
   \[\implies\] How well do we predict performance?
How to interpret the evaluation results?
Our proposal: linear regression

\[
\text{model performance} = \beta_0 + \beta_1 \cdot p_1 + \beta_2 \cdot p_2 + \beta_3 \cdot p_{1\ast2} + \ldots + \epsilon
\]

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?
How to interpret the evaluation results?
Our proposal: linear regression

$$\text{model performance} = \beta_0 + \beta_1 \cdot p_1 + \beta_2 \cdot p_2 + \beta_3 \cdot p_1 \cdot p_2 + \ldots + \epsilon$$

1. **Adjusted $R^2$:** proportion of variance explained by the model
   \(\rightsquigarrow\) How well do we predict performance?

2. **Feature ablation:** proportion of variance explained by a parameter together with all its interactions
   \(\rightsquigarrow\) Which parameters affect performance the most?

3. **Model predictions:** visualization of predicted performance
   \(\rightsquigarrow\) What are the best parameter values?
How well do we predict performance?
A concrete example: TOEFL, SVD (504k data points)

accuracy \sim ...
How well do we predict performance?
A concrete example: TOEFL, SVD (504k data points)

\[
\text{accuracy} \sim \text{corpus} + \text{window} + \text{score} + \text{transformation} \\
+ \text{metric} + \text{rel.index}
\]

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

Model fit: \( \text{Adj.R}^2 \)

| 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)

\[\text{accuracy} \sim \text{corpus} + \text{window} + \text{score} + \text{transformation} + \text{metric} + \text{rel.index} + n.\text{dim} + \text{dim.skip}\]

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

Model fit: \(\text{Adj.} R^2\)
- basic \(43\%\)
- \& SVD \(+24\%\)

Assumption: a good linear model acts as a “smoothing" algorithm which filters away random noise & captures robust trends.
Methodology for DSM Evaluation
Interpreting DSM performance with linear regression

How well do we predict performance?
A concrete example: TOEFL, SVD (504k data points)

\[
\text{accuracy} \sim \text{corpus} \times \text{window} \times \text{score} \times \text{transformation} \\
\times \text{metric} \times \text{rel.index} \times \text{n.dim} \times \text{dim.skip}
\]

<table>
<thead>
<tr>
<th>corpus</th>
<th>window</th>
<th>score</th>
<th>transformation</th>
<th>metric</th>
<th>n.dim</th>
<th>dim.skip</th>
<th>rel.index</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>wacky</td>
<td>8</td>
<td>t-score</td>
<td>none</td>
<td>manhattan</td>
<td>700</td>
<td>0</td>
<td>dist</td>
<td>71.25</td>
</tr>
<tr>
<td>bnc</td>
<td>16</td>
<td>z-score</td>
<td>root</td>
<td>cosine</td>
<td>100</td>
<td>100</td>
<td>rank</td>
<td>75.00</td>
</tr>
<tr>
<td>wacky</td>
<td>16</td>
<td>MI</td>
<td>log</td>
<td>cosine</td>
<td>100</td>
<td>50</td>
<td>dist</td>
<td>77.50</td>
</tr>
<tr>
<td>bnc</td>
<td>8</td>
<td>frequency</td>
<td>none</td>
<td>cosine</td>
<td>900</td>
<td>50</td>
<td>rank</td>
<td>75.00</td>
</tr>
<tr>
<td>ukwac</td>
<td>16</td>
<td>MI</td>
<td>none</td>
<td>cosine</td>
<td>500</td>
<td>100</td>
<td>rank</td>
<td>81.25</td>
</tr>
<tr>
<td>bnc</td>
<td>8</td>
<td>tf.idf</td>
<td>root</td>
<td>cosine</td>
<td>300</td>
<td>100</td>
<td>rank</td>
<td>75.00</td>
</tr>
<tr>
<td>bnc</td>
<td>16</td>
<td>tf.idf</td>
<td>root</td>
<td>manhattan</td>
<td>300</td>
<td>100</td>
<td>dist</td>
<td>51.25</td>
</tr>
<tr>
<td>ukwac</td>
<td>2</td>
<td>tf.idf</td>
<td>log</td>
<td>manhattan</td>
<td>300</td>
<td>50</td>
<td>rank</td>
<td>53.75</td>
</tr>
<tr>
<td>ukwac</td>
<td>1</td>
<td>simple-l1</td>
<td>log</td>
<td>manhattan</td>
<td>500</td>
<td>100</td>
<td>dist</td>
<td>85.00</td>
</tr>
</tbody>
</table>

Model fit: \(\text{Adj.R}^2\)
- 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

<table>
<thead>
<tr>
<th>Effect</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>score</td>
<td>10.53</td>
</tr>
<tr>
<td>score:transformation</td>
<td>7.42</td>
</tr>
<tr>
<td>score:metric</td>
<td>1.77</td>
</tr>
<tr>
<td>corpus:score</td>
<td>0.84</td>
</tr>
<tr>
<td>score:context.dim</td>
<td>0.64</td>
</tr>
<tr>
<td>other int. &lt; 0.5</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Feature ablation</strong></td>
<td><strong>22.13</strong></td>
</tr>
</tbody>
</table>
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

© Evert/Lenci/Baroni/Lapesa (CC-by-sa)
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)
Contrasting semantic relations (Lapesa et al. 2014)

Datasets: Semantic Priming Project, GEK priming dataset
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**
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
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 :)

© Evert/Lenci/Baroni/Lapesa (CC-by-sa)
Coming soon . . .

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

© Evert/Lenci/Baroni/Lapesa (CC-by-sa) DSM Tutorial – Part 3 wordspace.collocations.de 43 / 80
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

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

DS beyond NLP: Linguistic evaluation
  Polysemy
  Compositionality
  Non distributional meaning
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

▶ **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)

Target: bank

$\text{bank}_1$: The broker went to the bank to secure his cash
$\text{bank}_2$: 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

```r
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

<table>
<thead>
<tr>
<th></th>
<th>breed</th>
<th>tail</th>
<th>feed</th>
<th>kill</th>
<th>important</th>
<th>explain</th>
<th>likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>84</td>
<td>17</td>
<td>8</td>
<td>38</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>dog</td>
<td>579</td>
<td>14</td>
<td>32</td>
<td>63</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>animal</td>
<td>45</td>
<td>11</td>
<td>86</td>
<td>136</td>
<td>13</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>time</td>
<td>19</td>
<td>8</td>
<td>29</td>
<td>134</td>
<td>94</td>
<td>44</td>
<td>100</td>
</tr>
<tr>
<td>reason</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>18</td>
<td>71</td>
<td>140</td>
<td>39</td>
</tr>
<tr>
<td>cause</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>55</td>
<td>35</td>
<td>51</td>
</tr>
<tr>
<td>effect</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>62</td>
<td>37</td>
<td>14</td>
</tr>
</tbody>
</table>
```
Context vectors: can we do it in wordspace?
Yes :D

"cats and dogs need their time"

```r
> context.vectors(TT, s1)

breed  tail  feed  kill  important  explain  likely
1  227.3333  13  23  78.3333  31.6667  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"

```r
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...

```r
# 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
```

<table>
<thead>
<tr>
<th></th>
<th>breed</th>
<th>tail</th>
<th>feed</th>
<th>kill</th>
<th>important</th>
<th>explain</th>
<th>likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>84.00</td>
<td>17.00</td>
<td>8</td>
<td>38.00</td>
<td>0.00</td>
<td>2.00</td>
<td>0</td>
</tr>
<tr>
<td>dog</td>
<td>579.00</td>
<td>14.00</td>
<td>32</td>
<td>63.00</td>
<td>1.00</td>
<td>2.00</td>
<td>2</td>
</tr>
<tr>
<td>animal</td>
<td>45.00</td>
<td>11.00</td>
<td>86</td>
<td>136.00</td>
<td>13.00</td>
<td>5.00</td>
<td>4</td>
</tr>
<tr>
<td>time</td>
<td>19.00</td>
<td>8.00</td>
<td>29</td>
<td>134.00</td>
<td>94.00</td>
<td>44.00</td>
<td>100</td>
</tr>
<tr>
<td>reason</td>
<td>1.00</td>
<td>0.00</td>
<td>1</td>
<td>18.00</td>
<td>71.00</td>
<td>140.00</td>
<td>39</td>
</tr>
<tr>
<td>cause</td>
<td>0.00</td>
<td>1.00</td>
<td>0</td>
<td>3.00</td>
<td>55.00</td>
<td>35.00</td>
<td>51</td>
</tr>
<tr>
<td>effect</td>
<td>0.00</td>
<td>1.00</td>
<td>1</td>
<td>6.00</td>
<td>62.00</td>
<td>37.00</td>
<td>14</td>
</tr>
<tr>
<td>s1</td>
<td>227.33</td>
<td>13.00</td>
<td>23</td>
<td>78.33</td>
<td>31.67</td>
<td>16.00</td>
<td>34</td>
</tr>
<tr>
<td>s2</td>
<td>6.33</td>
<td>3.33</td>
<td>10</td>
<td>47.67</td>
<td>70.33</td>
<td>38.67</td>
<td>55</td>
</tr>
</tbody>
</table>
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...

```r
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))

dog
cat
animal
time
reason
effect
cause
s1
s2

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, l_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](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

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

DS beyond NLP: Linguistic evaluation
  Polysemy
  Compositionality
  Non distributional meaning
Compositionality
Can we capture it in DS?

- Formally: compositionality implies some operator $\oplus$ such that
  \[
  \text{meaning}(w_1 w_2) = \text{meaning}(w_1) \oplus \text{meaning}(w_2)
  \]
- CDSM recipe
  - Distributional vectors for meaning($w_1$) and meaning($w_2$)
  - Operators: mathematical strategies 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)

<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>problem</td>
<td>2</td>
<td>15</td>
<td>7</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>
Compositionality with distributional vectors

Additive and Multiplicative Models (Mitchell and Lapata, 2010)

<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>problem</td>
<td>2</td>
<td>15</td>
<td>7</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ p = u + v \]

\[ \text{predicted(practical, difficulty)} = \text{practical} + \text{difficulty} = [1 \ 14 \ 6 \ 14 \ 4] \]
Compositionality with distributional vectors

Additive and Multiplicative Models (Mitchell and Lapata, 2010)

<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>create</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>problem</td>
<td>2</td>
<td>15</td>
<td>7</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ p = u + v \]

predicted(practical difficulty) = \textbf{practical} + \textbf{difficulty} = [1 14 6 14 4]

\[ p = u \odot v \]

predicted(practical difficulty) = \textbf{practical} \odot \textbf{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 ⊗ difficulty?

2. **Quantitative evaluation**
   - Collect a vector for "practical difficulty" in (obviously the same) corpus: `observed(practical difficulty)`
   - `observed(practical difficulty) ≈ predicted(practical difficulty)`
     - Which of the two produces a better approximation?
       - practical + difficulty or practical ⊗ 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.)

<table>
<thead>
<tr>
<th>bad luck</th>
<th>electronic communities</th>
<th>historical map</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad</td>
<td>electronic storage</td>
<td>topographical</td>
</tr>
<tr>
<td>bad weekend</td>
<td>electronic transmission</td>
<td>atlas</td>
</tr>
<tr>
<td>good spirit</td>
<td>purpose</td>
<td>historical material</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>important route</td>
<td>nice girl</td>
<td>little war</td>
</tr>
<tr>
<td>important transport</td>
<td>good girl</td>
<td></td>
</tr>
<tr>
<td>important road</td>
<td>big girl</td>
<td>great war</td>
</tr>
<tr>
<td>major road</td>
<td>guy</td>
<td>major war</td>
</tr>
<tr>
<td>red cover</td>
<td>special collection</td>
<td>small war</td>
</tr>
<tr>
<td>black cover</td>
<td>general collection</td>
<td>small son</td>
</tr>
<tr>
<td>hardback</td>
<td>small collection</td>
<td>small daughter</td>
</tr>
<tr>
<td>red label</td>
<td>archives</td>
<td>mistress</td>
</tr>
</tbody>
</table>

- **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

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

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 ≥ 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)
   - 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)

1. **Input**: derived/stem vector pairs for each affix
   - un-: unfaithful/faithful, unbiased/biased, unwell/well
   - -ly: true/truly, mad/madly, deep/deeply
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)
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

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

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

DS beyond NLP: Linguistic evaluation
  Polysemy
  Compositionality

Non distributional meaning
Not all Semantic Knowledge is Distributional
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
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

Wrapping up
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
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
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 :) )

ителя 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

- AMTMT set
  - 402 nouns, 21 classes
  - day ⇒ TIME
  - kiwi ⇒ FRUIT
  - kitten ⇒ ANIMAL
  - volleyball ⇒ GAME

- SAT16 set
  - 83 nouns, 10 classes
  - chicken ⇒ BIRD
  - bear ⇒ LAND.MAMMAL
  - pot ⇒ KITCHENWARE
  - oak ⇒ TREE

- CSSLI 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


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

References III


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.

References V


References VI


References VII


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