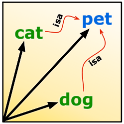
Hands-on Distributional Semantics

Part 5: DS beyond NLP – Free association norms

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

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

http://wordspace.collocations.de/doku.php/course:esslli2021:start



Outline

Cognitive modelling with DSM

- **Why?** – Because we want to know whether DS captures the mental lexical knowledge of human speakers!

Cognitive modelling with DSM

- **Why?** – Because we want to know whether DS captures the mental lexical knowledge of human speakers!
- Task: DSM predicts reaction times in **priming experiments** (Hare *et al.* 2009; Lapesa & Evert 2013)
  - often just experimental items used for multiple-choice task (e.g. Padó & Lapata 2007; Herdağdelen *et al.* 2009)
  - cf. tasks constructed from `Lazaridou2013` yesterday
  - data sets of experimental items: `GEK_Items`, `SPP_Items`

Cognitive modelling with DSM

- **Why?** – Because we want to know whether DS captures the mental lexical knowledge of human speakers!
- Task: DSM predicts reaction times in **priming experiments** (Hare *et al.* 2009; Lapesa & Evert 2013)
  - often just experimental items used for multiple-choice task (e.g. Padó & Lapata 2007; Herdağdelen *et al.* 2009)
  - cf. tasks constructed from `Lazaridou2013` yesterday
  - data sets of experimental items: `GEK_Items`, `SPP_Items`
- Task: DSM predicts **EEG potentials** (Murphy *et al.* 2009) or **fMRI brain activation** levels (Mitchell *et al.* 2008)
  - huge datasets, but tiny and selective vocabulary

Cognitive modelling with DSM

- **Why?** – Because we want to know whether DS captures the mental lexical knowledge of human speakers!
- Task: DSM predicts reaction times in **priming experiments** (Hare *et al.* 2009; Lapesa & Evert 2013)
  - often just experimental items used for multiple-choice task (e.g. Padó & Lapata 2007; Herdağdelen *et al.* 2009)
  - cf. tasks constructed from `Lazaridou2013` yesterday
  - data sets of experimental items: `GEK_Items`, `SPP_Items`
- Task: DSM predicts **EEG potentials** (Murphy *et al.* 2009) or **fMRI brain activation** levels (Mitchell *et al.* 2008)
  - huge datasets, but tiny and selective vocabulary
- Task: DSM predicts human **free associations**
  - often considered a "window into the mental lexicon"
  - free association norms available for thousands of cue words

Free associations

... a cue into the organization of the mental lexicon?

Which words come to your mind if you hear ...

- whisky $\rightarrow$

Free associations

... a cue into the organization of the mental lexicon?

Which words come to your mind if you hear ...

- whisky → gin, drink, scotch, bottle, soda
- giraffe →

Free associations

... a cue into the organization of the mental lexicon?

Which words come to your mind if you hear ...

- whisky → gin, drink, scotch, bottle, soda
- giraffe → neck, animal, zoo, long, tall

Free associations

... a cue into the organization of the mental lexicon?

Which words come to your mind if you hear ...

- whisky → gin, drink, scotch, bottle, soda
- giraffe → neck, animal, zoo, long, tall

- Hypotheses concerning the nature of the underlying process:
  - Result of learning-by-contiguity (James 1890)
    ☞ syntagmatic (1st-order)
  - Result of symbolic processes which make use of complex semantic structures (Clark 1970) ☞ paradigmatic (2nd-order)

Free associations

... a cue into the organization of the mental lexicon?

Which words come to your mind if you hear ...

- whisky → gin, drink, scotch, bottle, soda
- giraffe → neck, animal, zoo, long, tall

- Hypotheses concerning the nature of the underlying process:
  - Result of learning-by-contiguity (James 1890)
    ☞ syntagmatic (1st-order)
  - Result of symbolic processes which make use of complex semantic structures (Clark 1970) ☞ paradigmatic (2nd-order)
- Large collections available
  - Edinburgh Associative Thesaurus (**EAT**)
    8210 stimuli, 100 subjects (Kiss *et al.* 1973)
  - University of South Florida Free Association Norms (**USF**)
    5019 stimuli, 6000 subjects (Nelson *et al.* 2004)

Free associations in a DSM

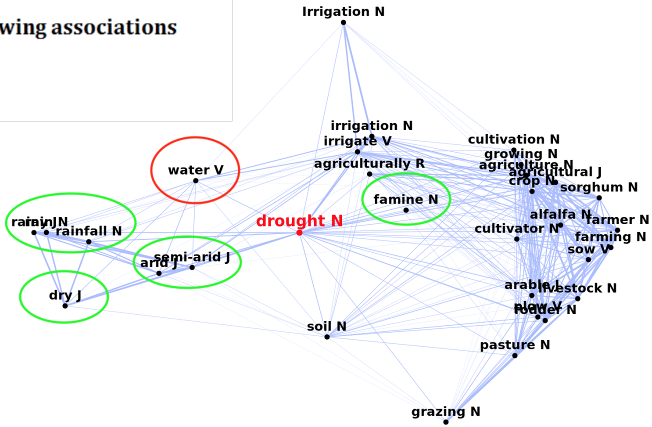
Drought in EAT vs. DSM

drought stimulated the following associations

Number of different answers: 30

Total count of all answers: 97

- WATER 21 0.22
- DRY 16 0.16
- THIRST 9 0.09
- FAMINE 7 0.07
- RAIN 7 0.07
- DESERT 6 0.06
- BEER 5 0.05
- CRACK 2 0.02
- HOT 2 0.02
- SAND 2 0.02
- ALE 1 0.01
- ARID 1 0.01
- AUSTRALIA 1 0.01
- CATTLE 1 0.01
- COLD 1 0.01
- COOL 1 0.01
- DEATH 1 0.01
- DUST 1 0.01
- GALE 1 0.01
- MONSOON 1 0.01



Free associations & co-occurrence data

Previous work

- Wettler *et al.* (2005)
  - Data: subset of EAT (100 stimuli)
  - Task: prediction of the most common free associate
  - Model: first-order model, BNC, large window (20 words)
  - Result: human associative responses can be predicted from contiguities between words in language use (collocations)
- ESSLLI 2008 Shared Task
  - Data: subset of EAT (a different set of 100 stimuli)
  - Task 1: discrimination btw. the most common associate and hapax/random distractors ➜ multiple choice
  - Task 2: prediction of the most common free associate
  - Result: first-order models (collocations) are better than second-order models (DSMs)

Outline

Problems of standard tasks & data sets

Problems with semantic interpretation of DSMs don't only stem from evaluation methodology . . .

. . . **data sets can be problematic as well!**

Problems of standard tasks & data sets

Problems with semantic interpretation of DSMs don't only stem from evaluation methodology . . .

. . . **data sets can be problematic as well!**

Two major problems:

- DSMs may exploit contingent properties of the task
  - random fillers as distractors ("controls")
    ➥ recognize random word pairs rather than semantic relations
  - choice of clearly separated categories and prototypical exemplars in noun clustering task (ESSLLI 2008)
    ➥ much harder to identify categories in general word list
  - typical superordinate-level words in hypernym detection task
    ➥ recognize "typical hypernym" in a multiple-choice setting
- Data set size too small
  - e.g. 97.5% accuracy on 80 TOEFL items ➜ over-fitting

DSM evaluation problems: a concrete example

The CogALex-V Shared Task (Santus *et al.* 2016)

- Aim: better linguistic understanding of DS from identification of specific **semantic relations**
- Data: 747 target words with approx. 10 candidate relata each
  - training set: 318 targets, 3054 word pairs
  - test set: 429 targets, 4260 word pairs

DSM evaluation problems: a concrete example

The CogALex-V Shared Task (Santus *et al.* 2016)

- Aim: better linguistic understanding of DS from identification of specific **semantic relations**
- Data: 747 target words with approx. 10 candidate relata each
  - training set: 318 targets, 3054 word pairs
  - test set: 429 targets, 4260 word pairs
- Subtask 1: related vs. unrelated word pairs
  - unrelated pairs are random fillers
  - ☞ relatively easy: $F_1 = 79.0\%$ (best system)

DSM evaluation problems: a concrete example

The CogALex-V Shared Task (Santus *et al.* 2016)

- Aim: better linguistic understanding of DS from identification of specific **semantic relations**
- Data: 747 target words with approx. 10 candidate relata each
  - training set: 318 targets, 3054 word pairs
  - test set: 429 targets, 4260 word pairs
- Subtask 1: related vs. unrelated word pairs
  - unrelated pairs are random fillers
  - ☞ relatively easy: $F_1 = 79.0\%$ (best system)
- Subtask 2: distinguish between semantic relations
  - SYN: w_2 can be used with same meaning as w_1
  - ANT: w_2 can be used as the opposite of w_1
  - HYPER: w_1 is a kind of w_2
  - PART_OF: w_1 is a part of w_2
  - RANDOM: no relation (random word + manual check)
  - ☞ relatively hard: $F_1 = 44.5\%$ (best system: deep learning)

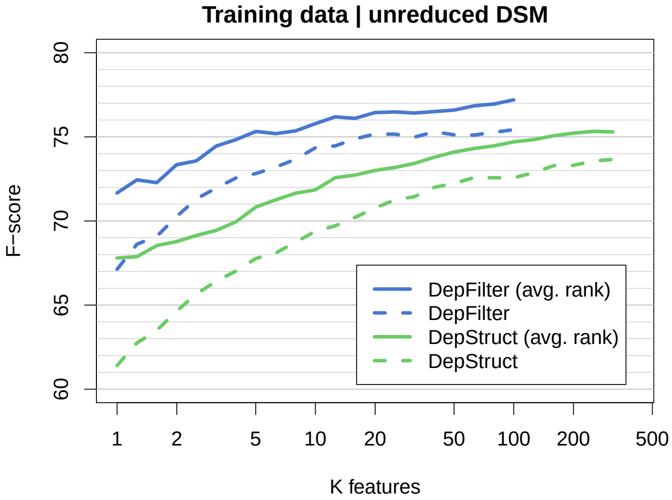
DSM evaluation problems: a concrete example

Mach 5 at CogALex 2016 (Evert 2016)

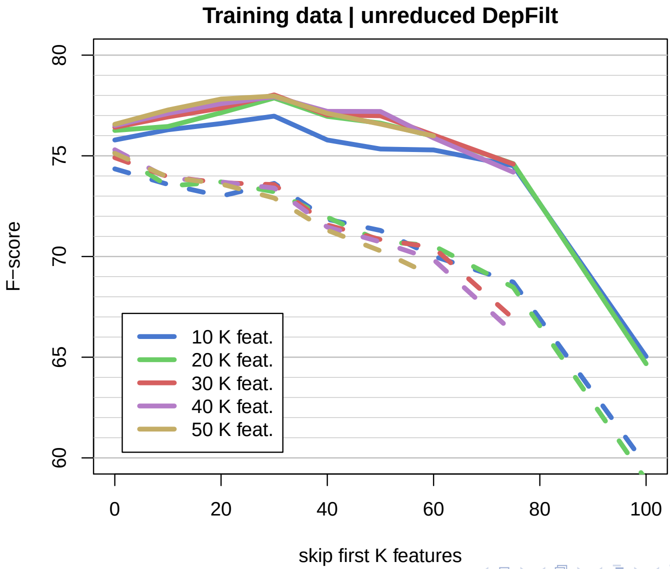
- Mach 5 participated in the CogALex-V Shared Task as a traditional "count" (non-neural) DSM
  - 10-billion-word Web corpus (Schäfer & Bildhauer 2012)
  - syntactic dependencies from C&C parser (Curran *et al.* 2007)
  - 26.5k target words, up to 300k feature dimensions
  - other parameters set according to Lapesa & Evert (2014)
- Parameter optimization on training data (subtask 1)
- Machine learning on optimized representations (subtask 2)
  - learns relevance weights for 600 latent SVD dimensions
  - best results from combination of different SVD spaces

☞ Try it yourself: `http://www.collocations.de/data/#mach5`

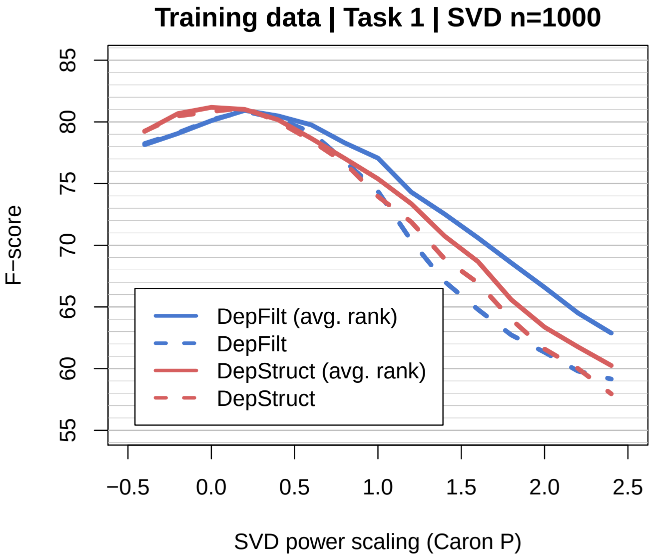
Mach 5: Parameter optimization



Mach 5: Parameter optimization



Mach 5: Parameter optimization



Mach 5: Are we doing well?

$F_1 = 77.88\%$ for related vs. unrelated (best: 79.0%)

However . . .

Mach 5: Are we doing well?

$F_1 = 77.88\%$ for related vs. unrelated (best: 79.0%)

However . . .

- Parameter optimization yields surprising result: best model uses < 50k features with relatively low frequency

Mach 5: Are we doing well?

$F_1 = 77.88\%$ for related vs. unrelated (best: 79.0%)

However ...

- Parameter optimization yields surprising result: best model uses < 50k features with relatively low frequency
- Nearest neighbours are unsatisfactory, e.g. for *play*: *playing* (54.1°), *star* (62.8°), *reunite* (62.9°), *co-star* (64.3°), *reprise* (64.4°), *player* (66.7°), *score* (68.5°), *audition* (69.2°), *sing* (69.4°), *actor* (69.5), *understudy* (69.6), *game* (70.3), ...

Mach 5: Are we doing well?

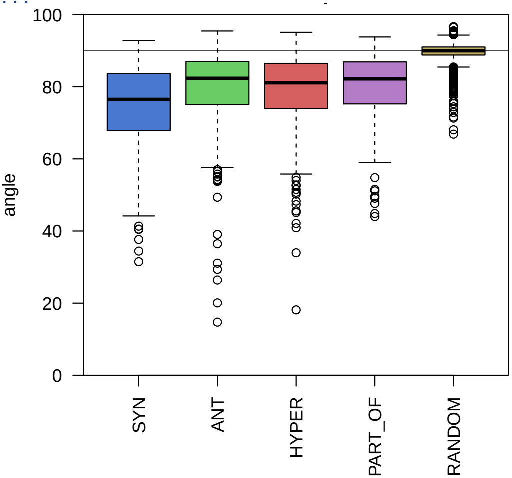
$F_1 = 77.88\%$ for related vs. unrelated (best: 79.0%)

However ...

- Parameter optimization yields surprising result:
  best model uses < 50k features with relatively low frequency
- Nearest neighbours are unsatisfactory, e.g. for *play*:
  playing (54.1°), *star* (62.8°), *reunite* (62.9°), *co-star* (64.3°), *reprise* (64.4°), *player* (66.7°), *score* (68.5°), *audition* (69.2°), *sing* (69.4°), *actor* (69.5), *understudy* (69.6), *game* (70.3), ...
- Why is Mach 5 still doing so well in the task, then?

Mach 5: What is going wrong?

A disturbing result . . .



- ☞ DSM has learned to recognize random word pairs (at 90°)!
- ☞ We need better data sets with **high-quality distractors**!

Outline

The Free ASsociation Task (FAST) data set

Preprocessing

1. Starting point: EAT (8210 stimuli), USF (5019 stimuli)

The Free ASsociation Task (FAST) data set

Preprocessing

1. Starting point: EAT (8210 stimuli), USF (5019 stimuli)

2. Out-of-context POS tagging
   - Annotate items in EAT and USF (stimuli and responses) with part of speech information
   - How? Most frequent POS in Web corpus ENCOW: publicly available 10-billion-word Web corpus ➜ replicability

The Free ASsociation Task (FAST) data set

Preprocessing

1. Starting point: EAT (8210 stimuli), USF (5019 stimuli)

2. Out-of-context POS tagging
   - Annotate items in EAT and USF (stimuli and responses) with part of speech information
   - How? Most frequent POS in Web corpus ENCOW: publicly available 10-billion-word Web corpus ➜ replicability

3. Out-of-context lemmatization
   - `morpha`, a robust morphological analyzer `http://users.sussex.ac.uk/~johnca/morph.html`
   - lemmatization of unknown words based on POS tag

The Free ASsociation Task (FAST) data set

Preprocessing

1. Starting point: EAT (8210 stimuli), USF (5019 stimuli)

2. Out-of-context POS tagging
   - Annotate items in EAT and USF (stimuli and responses) with part of speech information
   - How? Most frequent POS in Web corpus ENCOW: publicly available 10-billion-word Web corpus ➜ replicability

3. Out-of-context lemmatization
   - `morpha`, a robust morphological analyzer `http://users.sussex.ac.uk/~johnca/morph.html`
   - lemmatization of unknown words based on POS tag

4. Annotation with frequency information
   - frequency lists from ENCOW (lemmatised with `morpha`)

The Free ASsociation Task (FAST) data set

Item selection

For each stimulus in EAT (8210) and USF (5019) select a:

(multiwords, numbers, closed-class words, and other words that do not occur in ENCOW were discarded)

The Free ASsociation Task (FAST) data set

Item selection

For each stimulus in EAT (8210) and USF (5019) select a:

- FIRST: the most common associate response

(multiwords, numbers, closed-class words, and other words that do not occur in ENCOW were discarded)

The Free ASsociation Task (FAST) data set

Item selection

For each stimulus in EAT (8210) and USF (5019) select a:

- FIRST: the most common associate response
- HAPAX: a response generated for the target once
  - or twice for USF (hapax responses are omitted there)
  - if several HAPAX candidates are available, pick the one whose lemma frequency matches most closely that of FIRST

(multiwords, numbers, closed-class words, and other words that do not occur in ENCOW were discarded)

The Free ASsociation Task (FAST) data set

Item selection

For each stimulus in EAT (8210) and USF (5019) select a:

- FIRST: the most common associate response
- HAPAX: a response generated for the target once
  - or twice for USF (hapax responses are omitted there)
  - if several HAPAX candidates are available, pick the one whose lemma frequency matches most closely that of FIRST
- RANDOM, by randomly picking a word which was among the top 25% associates *of another stimulus* (and produced at least 5 times). If possible:
  - match lemma frequency of RANDOM and FIRST
  - try to use each RANDOM only once

(multiwords, numbers, closed-class words, and other words that do not occur in ENCOW were discarded)

The FAST data set

Final data set

- EAT subset: **3836** test items + **3774** training items
- USF subset: **2359** test items + **2360** training items
- Item = (STIMULUS, FIRST, HAPAX, RANDOM)

- Each stimulus and candidate response provided as lowercased word form and POS-disambiguated lemma
  - + ENCOW frequency information
  - + # test subjects who produced response

- Download: `https://osf.io/cd8ar/` (Evert & Lapesa 2021)
- Included as `FAST` in package `wordspaceEval`

The FAST dataset

The new EAT task isn't perfect either . . . yet

- Guessing POS from corpus doesn't always work
  - e.g. *fit*$_{\text{VERB}}$ → *epileptic*$_{\text{ADJ}}$, *aristocracy*$_{\text{NOUN}}$ → *lords*$_{\text{NAME}}$
  - but very few lemmatization errors (e.g. *daiquiri* → *daiquirus*)

The FAST dataset

The new EAT task isn't perfect either . . . yet

- Guessing POS from corpus doesn't always work
  - e.g. *fit*$_{\text{VERB}}$ → *epileptic*$_{\text{ADJ}}$, *aristocracy*$_{\text{NOUN}}$ → *lords*$_{\text{NAME}}$
  - but very few lemmatization errors (e.g. *daiquiri* → *daiquirus*)
- Colloquialisms and British slang
  - e.g. *bod*$_{\text{NOUN}}$ → *person*$_{\text{NOUN}}$ (rare in written corpus)
  - but Web corpus has Welsh *bod* 'to be' mistagged as noun
  - DSM neighbours: *yn*, *hynny*, *mewn*, *hwn*, *gyfer*, . . . , 49. *bloke*, *techy*$_{\text{NOUN}}$, *nus*, *hon*, . . . , 60. *guy*, *mai*, *geezer*, . . .
  - another example is *mellow*$_{\text{ADJ}}$ → *yellow*$_{\text{ADJ}}$

The FAST tasks

Task 1: **multiple-choice**

- Given a stimulus and a <FIRST, HAPAX, RANDOM> triple, determine which of the three candidates is FIRST.
  - Stimulus: *accept*, < <u>*receive*</u>, *love*, *soul*>
- **Performance**: accuracy
- **Baseline**: 33.3%

The FAST tasks

Task 2: **open-vocabulary lexical access**

- Given a stimulus (e.g., *accept*), predict FIRST (*receive*) out of a candidate set (all FIRST: USF=1197, EAT=1633)
- **Performance**: two measures
  - **Soft accuracy**: average over reciprocal rank ($1/r$) of the true FIRST associate, as a percentage.
    - similar to accuracy of predicting first associate, but awards partial points for almost correct guesses
    - always $\geq$ top-1 accuracy
  - **Log rank**: geometric mean of r across all stimuli.
    - corresponds to average over $\log r$
    - better differentiation for models that rarely get the correct answer (and hence score low on soft accuracy)
- **Baselines**
  - **Soft accuracy**: USF=0.64% and EAT=0.49%
  - **Log rank**: USF=442.0 and EAT=602.4

Outline

Experimental setup

- **DSMs (second-order)**: symmetric span of 2 vs. 10 words, other parameters set according to Lapesa & Evert (2014).
  - we experiment with Caron P (Bullinaria & Levy 2012)
  - $P = 0$ equalizes contributions of SVD dimensions

Experimental setup

- **DSMs (second-order)**: symmetric span of 2 vs. 10 words, other parameters set according to Lapesa & Evert (2014).
  - we experiment with Caron P (Bullinaria & Levy 2012)
  - $P = 0$ equalizes contributions of SVD dimensions
- **Collocations (first-order)**: symmetric span, 2 vs. 10 words, with four different association measures (Evert 2008)
  - conditional probability $P(w_2|w_1)$
  - log-likelihood $\log G^2$ (popular for collocations)
  - $\mathrm{MI}^2 = \log_2 \frac{O^2}{E}$ = geometric mean of $P(w_2|w_1)$ and $P(w_1|w_2)$
  - PPMI (popular for DSMs)

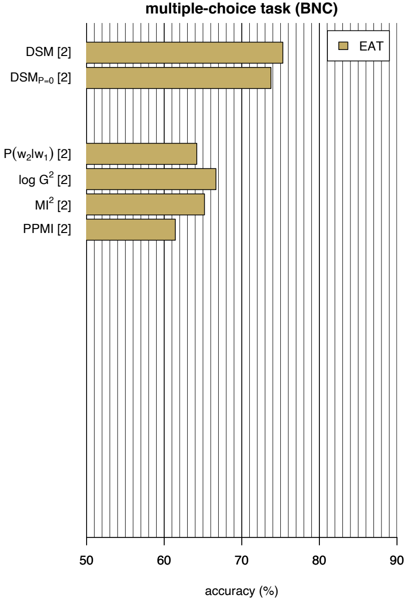
Experimental setup

- **DSMs (second-order)**: symmetric span of 2 vs. 10 words, other parameters set according to Lapesa & Evert (2014).
  - we experiment with Caron P (Bullinaria & Levy 2012)
  - $P = 0$ equalizes contributions of SVD dimensions
- **Collocations (first-order)**: symmetric span, 2 vs. 10 words, with four different association measures (Evert 2008)
  - conditional probability $P(w_2|w_1)$
  - log-likelihood $\log G^2$ (popular for collocations)
  - $\mathrm{MI}^2 = \log_2 \frac{O^2}{E}$ = geometric mean of $P(w_2|w_1)$ and $P(w_1|w_2)$
  - PPMI (popular for DSMs)
- **Corpus data**: for DSMs and collocations
  - British National Corpus: 100M words
  - ENCOW 2014 Web corpus, unique sentences: 8.5G words

Experimental setup

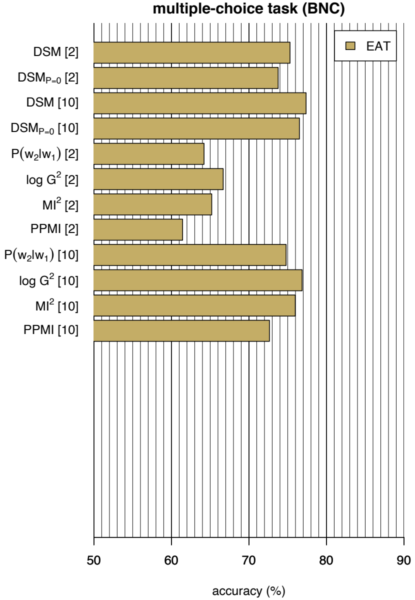
- **DSMs (second-order)**: symmetric span of 2 vs. 10 words, other parameters set according to Lapesa & Evert (2014).
  - we experiment with Caron P (Bullinaria & Levy 2012)
  - $P = 0$ equalizes contributions of SVD dimensions
- **Collocations (first-order)**: symmetric span, 2 vs. 10 words, with four different association measures (Evert 2008)
  - conditional probability $P(w_2|w_1)$
  - log-likelihood $\log G^2$ (popular for collocations)
  - $\mathrm{MI}^2 = \log_2 \frac{O^2}{E}$ = geometric mean of $P(w_2|w_1)$ and $P(w_1|w_2)$
  - PPMI (popular for DSMs)
- **Corpus data**: for DSMs and collocations
  - British National Corpus: 100M words
  - ENCOW 2014 Web corpus, unique sentences: 8.5G words
- **Neural embeddings**: pre-trained models
  - word2vec (Mikolov *et al.* 2013): 100G tokens of Google News
  - GloVe (Pennington *et al.* 2014): 6G tokens Wikipeda + Gigaword
  - GloVe: 42G tokens Web data (Common Crawl)
  - FastText (Joulin *et al.* 2017): 600G tokens Common Crawl

Results: Multiple-choice task



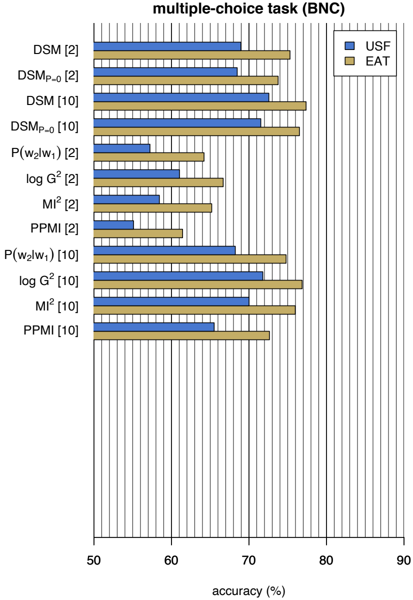
- British National Corpus (100M words)
- EAT subset

Results: Multiple-choice task



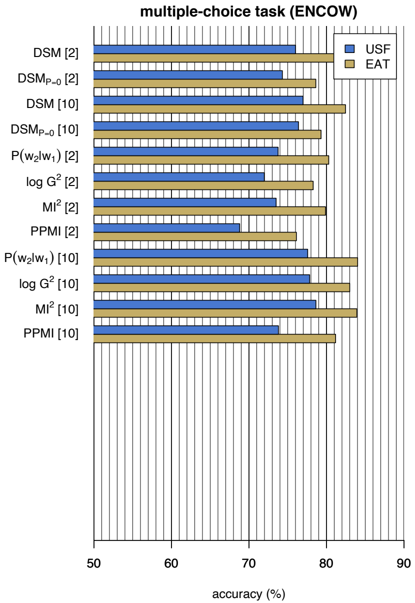
- British National Corpus (100M words)
- EAT subset

Results: Multiple-choice task



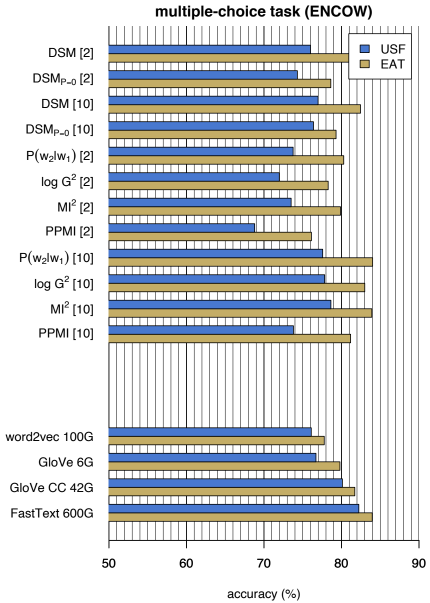
- British National Corpus (100M words)
- EAT vs. USF

Results: Multiple-choice task



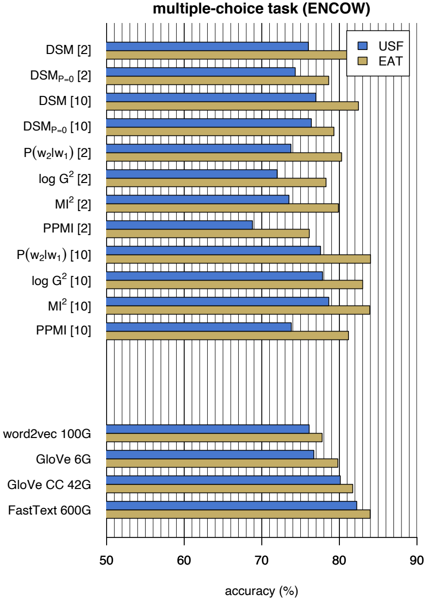
- ENCOW 2014 Web (8.5G words)
- EAT vs. USF

Results: Multiple-choice task



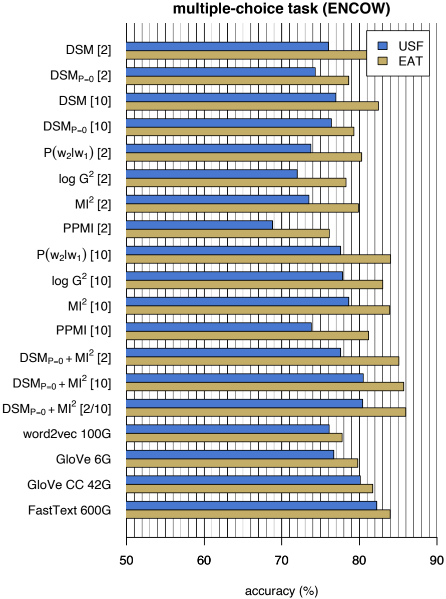
- ENCOW 2014 Web (8.5G words)
- EAT vs. USF
- Embeddings trained on much larger corpora

Results: Multiple-choice task



- ENCOW 2014 Web (8.5G words)
- EAT vs. USF
- Embeddings trained on much larger corpora
- Combined 1st-/2nd-order
  - $\mathrm{DSM}_{P=0} + \mathrm{MI}^2$
  - using neighbour rank
  - harmonic mean

Results: Multiple-choice task



- ENCOW 2014 Web (8.5G words)
- EAT vs. USF
- Embeddings trained on much larger corpora
- Combined 1st-/2nd-order
  - DSM$_{P=0}$ + MI^2
  - using neighbour rank
  - harmonic mean
  - competitive with state-of-the-art embeddings

Results: Multiple-choice task

model	span	$n = 2359$ USF	$n = 3836$ EAT
DSM	2	76.01%	81.78%
$\mathrm{DSM}_{P=0}$	2	74.31%	78.62%
DSM	10	76.98%	82.46%
$\mathrm{DSM}_{P=0}$	10	76.39%	79.30%
$P(w_2 \mid w_1)$	10	77.58%	84.02%
$\log G^2$	10	77.83%	83.00%
MI^2	2	78.64%	83.92%
PPMI	10	73.80%	81.18%
Combined	2	77.58%	85.09%
Combined	10	80.50%	85.71%
Combined	mix	80.41%	85.97%
word2vec	–	76.11%	77.78%
GloVe	–	76.71%	79.80%
GloVe CC	–	80.12%	81.72%
FastText	–	82.24%	83.97%

- ENCOW 2014 Web (8.5G words)
- EAT vs. USF
- Embeddings trained on much larger corpora
- Combined 1st-/2nd-order
  - $\mathrm{DSM}_{P=0} + \mathrm{MI}^2$
  - using neighbour rank
  - harmonic mean
  - competitive with state-of-the-art embeddings

Results: Open-choice task

		$n = 2359$ USF		$n = 3836$ EAT	
model	span	soft acc.	lrank	soft acc.	lrank
DSM	2	41.54%	6.6	34.53%	9.9
$\mathrm{DSM}_{P=0}$	2	42.12%	7.6	34.67%	12.1
DSM	10	42.01%	6.0	35.93%	9.1
$\mathrm{DSM}_{P=0}$	10	42.86%	7.1	35.68%	11.6

Results: Open-choice task

		$n = 2359$ USF		$n = 3836$ EAT	
model	span	soft acc.	lrank	soft acc.	lrank
DSM	2	41.54%	6.6	34.53%	9.9
$\text{DSM}_{P=0}$	2	42.12%	7.6	34.67%	12.1
DSM	10	42.01%	6.0	35.93%	9.1
$\text{DSM}_{P=0}$	10	42.86%	7.1	35.68%	11.6
$P(w_2 \mid w_1)$	10	22.34%	17.0	11.27%	27.1
$\log G^2$	10	37.63%	6.6	34.13%	8.8
MI^2	10	39.73%	6.2	34.01%	8.7
PPMI	10	35.34%	8.2	29.29%	12.2

Results: Open-choice task

		$n = 2359$ USF		$n = 3836$ EAT	
model	span	soft acc.	lrank	soft acc.	lrank
DSM	2	41.54%	6.6	34.53%	9.9
$\text{DSM}_{P=0}$	2	42.12%	7.6	34.67%	12.1
DSM	10	42.01%	6.0	35.93%	9.1
$\text{DSM}_{P=0}$	10	42.86%	7.1	35.68%	11.6
$P(w_2 \vert w_1)$	10	22.34%	17.0	11.27%	27.1
$\log G^2$	10	37.63%	6.6	34.13%	8.8
MI^2	10	39.73%	6.2	34.01%	8.7
PPMI	10	35.34%	8.2	29.29%	12.2
Combined	2	42.29%	5.5	37.54%	7.0
Combined	10	44.99%	4.8	39.48%	6.5
Combined	mix	45.36%	4.8	39.48%	6.4

Results: Open-choice task

model	span	USF ($n = 2359$) soft acc.	USF lrank	EAT ($n = 3836$) soft acc.	EAT lrank
DSM	2	41.54%	6.6	34.53%	9.9
$\text{DSM}_{P=0}$	2	42.12%	7.6	34.67%	12.1
DSM	10	42.01%	6.0	35.93%	9.1
$\text{DSM}_{P=0}$	10	42.86%	7.1	35.68%	11.6
$P(w_2 \mid w_1)$	10	22.34%	17.0	11.27%	27.1
$\log G^2$	10	37.63%	6.6	34.13%	8.8
MI^2	10	39.73%	6.2	34.01%	8.7
PPMI	10	35.34%	8.2	29.29%	12.2
Combined	2	42.29%	5.5	37.54%	7.0
Combined	10	44.99%	4.8	39.48%	6.5
Combined	mix	45.36%	4.8	39.48%	6.4
word2vec	–	38.98%	7.7	30.51%	14.8
GloVe	–	39.22%	7.6	30.19%	13.8
GloVe CC	–	44.01%	5.7	34.26%	10.5
FastText	–	51.00%	4.1	40.34%	7.2

Outline

Hands-on exercise

- Solve the FAST multiple-choice task with a DSM
  - `eval.multiple.choice()` does most of the work for you
  - use `details=TRUE` to inspect biggest mistakes and explore performance (e.g. wrt. frequency of stimulus and response)
- Can you also make use of first-order (collocation) data?
  - hint: the DSM matrix **M** contains co-occurrence counts
- Advanced: Can you combine DSMs with first-order data?
  - hint: use average of DSM and first-order "neighbour" rank
- Advanced: Try to solve the open-choice lexical access task
  - no ready-made evaluation function in `wordspace` yet
- R code in `hands_on_day5.R` will help you get started!

Bonus task: Reverse free associations

The CogALex-IV shared task (Rapp & Zock 2014)

Reverse multiword free association

- wheel, driver, bus, drive, lorry → ?
- away, minded, gone, present, ill → ?

- Data: subset of EAT (2000 stimuli each training/test)

Bonus task: Reverse free associations

The CogALex-IV shared task (Rapp & Zock 2014)

Reverse multiword free association

- wheel, driver, bus, drive, lorry → ?
- away, minded, gone, present, ill → ?

- Data: subset of EAT (2000 stimuli each training/test)
- Very challenging (best: 35% accuracy)
  - open-ended vocabulary (including inflected surface forms!)
  - need for integrating predictions of different stimuli
- And the winner was …
  - a system using first-order statistics to re-rank the output of a “standard" DSM (Ghosh *et al.* 2015)
  - our submission: best 1st-order: 27.7% / best 2nd-order: 14.0%
- Try it yourself: `CogALex4.rda`

Outline

Dimensionality reduction as matrix factorization

- PCA is based on **singular value decomposition** (**SVD**), which factorises any matrix $\mathbf{M}$ into

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\boldsymbol{\Sigma}$ is a diagonal matrix of **singular values** $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m > 0$

$$\begin{array}{cc} & n \\ k & \begin{bmatrix} & & \\ & \mathbf{M} & \\ & & \end{bmatrix} \end{array} = \begin{array}{cc} & m \\ k & \begin{bmatrix} & & \\ & \mathbf{U} & \\ & & \end{bmatrix} \end{array} \cdot \begin{array}{cc} & m \\ m & \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & \boldsymbol{\Sigma} & \sigma_m \end{bmatrix} \end{array} \cdot \begin{array}{cc} & n \\ m & \begin{bmatrix} & & \\ & \mathbf{V}^T & \\ & & \end{bmatrix} \end{array}$$

Dimensionality reduction as matrix factorization

- Columns $\mathbf{a}_i$ of $\mathbf{U}$ and $\mathbf{b}_i$ of $\mathbf{V}$ (**singular vectors**) are orthogonal ($\mathbf{a}_i^T \mathbf{a}_j = 0$) and of unit length ($\|\mathbf{a}_i\| = 1$)
- Key property: **truncated SVD** gives best least-squares approximation in r-dimensional subspace

$$\mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T = \begin{bmatrix} \vdots & & \vdots \\ \vdots & & \vdots \\ \mathbf{a}_1 & \cdots & \mathbf{a}_r \\ \vdots & & \vdots \\ \vdots & \mathbf{U}_r & \vdots \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & \mathbf{\Sigma}_r & \sigma_r \end{bmatrix} \cdot \begin{bmatrix} \cdots & \cdots & \mathbf{b}_1 & \cdots & \cdots \\ \mathbf{V}_r^T & & \vdots & & \\ \cdots & \cdots & \mathbf{b}_r & \cdots & \cdots \end{bmatrix}$$

Dimensionality reduction as matrix factorization

- Truncated SVD as orthogonal projection

$$\mathbf{M}\mathbf{V}_r = \mathbf{U}_r\mathbf{\Sigma}_r = \begin{bmatrix} \vdots & & \vdots \\ \sigma_1\mathbf{a}_1 & \cdots & \sigma_r\mathbf{a}_r \\ \vdots & & \vdots \end{bmatrix}$$

  ➜ `method="svd"` in `dsm.projection()`

- $\sigma_1^2 \geq \sigma_2^2 \geq \ldots =$ amount of distance information (i.e. variance of $\mathbf{M}$) captured by each **latent dimension**

☞ Catch up on the mathematics with Deisenroth *et al.* (2020)

Other matrix factorization techniques

- **Non-negative matrix factorization (NMF)**
  - **U** and **V** are stochastic matrices ($\mathbf{a}_i \geq 0$ and $\|\mathbf{a}_i\|_1 = 1$)
  - but no orthogonality constraints
  - cross-entropy instead of least-squares approximation
  - iterative algorithm with random initialisation for optimal rank-r approximation ($\neq$ sequence of ordered components)
  - see Lee & Seung (2001) and Boutsidis & Gallopoulos (2008)

Other matrix factorization techniques

- **Non-negative matrix factorization** (**NMF**)
  - **U** and **V** are stochastic matrices ($\mathbf{a}_i \geq 0$ and $\|\mathbf{a}_i\|_1 = 1$)
  - but no orthogonality constraints
  - cross-entropy instead of least-squares approximation
  - iterative algorithm with random initialisation for optimal rank-r approximation ($\neq$ sequence of ordered components)
  - see Lee & Seung (2001) and Boutsidis & Gallopoulos (2008)

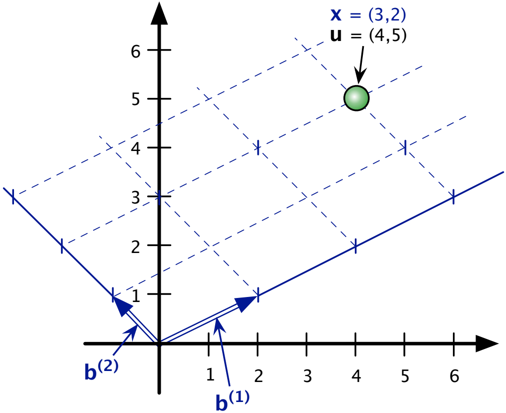
- NMF of term-document matrix $\Longleftrightarrow$ LDA **topic model**

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sigma_1 \mathbf{a}_1 \mathbf{b}_1^T + \sigma_2 \mathbf{a}_2 \mathbf{b}_2^T + \sigma_3 \mathbf{a}_3 \mathbf{b}_3^T + \dots$$

  - $\mathbf{a}_i$ = probability distribution of words in i-th topic
  - $\mathbf{b}_i$ = distribution of topic across documents

Other matrix factorization techniques

- NMF can be seen as **non-orthogonal** projection:
  $\mathbf{U\Sigma}$ = coordinates of projected points wrt. basis $\mathbf{V}$



Other matrix factorization techniques

- Levy *et al.* (2015, 213) show that **word2vec** embeddings implicitly factorize a shifted PPMI matrix
  - sigmoid loss function, weighted towards high frequencies
  - similarly, **GloVe** (Pennington *et al.* 2014) factorizes matrix of conditional probabilities with a frequency-weighted least-squares approximation

Other matrix factorization techniques

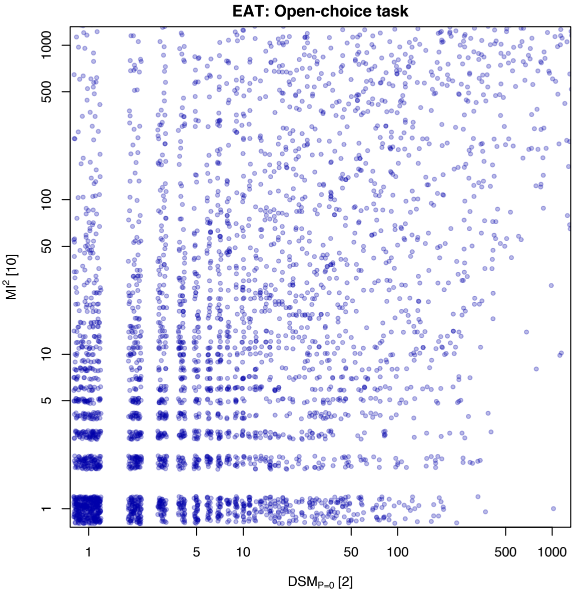
- Levy *et al.* (2015, 213) show that **word2vec** embeddings implicitly factorize a shifted PPMI matrix
  - sigmoid loss function, weighted towards high frequencies
  - similarly, **GloVe** (Pennington *et al.* 2014) factorizes matrix of conditional probabilities with a frequency-weighted least-squares approximation

- Explore matrix factorization techniques
  ☞ `hands_on_day5_matrix_factorization.R`

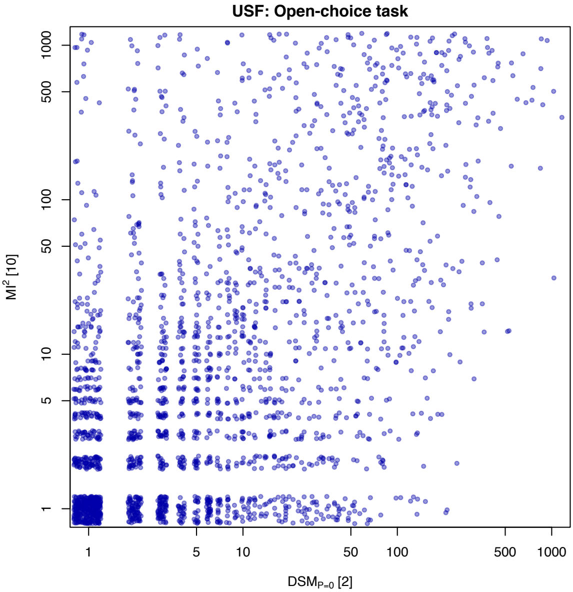
Outline

Syntagmatic vs. paradigmatic

Syntagmatic vs. paradigmatic



Syntagmatic vs. paradigmatic



Syntagmatic vs. paradigmatic

1^{st}-order = syntagmatic vs. 2^{nd}-order = paradigmatic?

- 1^{st}- and 2^{nd}-order models less complementary than expected
  ➥ relatively small benefit from combination
- But intuition not completely wrong (L2/R2):
  - DSM: *duckling* → *piglet, chick, duck, cygnet, hatchling, ...*
  - MI^2: *duckling* → *ugly, chick, duck, swan, fluffy, roast, ...*

Syntagmatic vs. paradigmatic

1st-order = syntagmatic vs. 2nd-order = paradigmatic?

- 1st- and 2nd-order models less complementary than expected
  ➥ relatively small benefit from combination
- But intuition not completely wrong (L2/R2):
  - DSM: *duckling* → *piglet, chick, duck, cygnet, hatchling, . . .*
  - MI^2: *duckling* → *ugly, chick, duck, swan, fluffy, roast, . . .*

Possible explanation for the overlap under (many) simplifying assumptions (sentence span, raw cooc freqs, . . .)

- Consider a term-context matrix **F** with very small contexts
  - e.g. **tweets**, sentences, paragraphs
  - or aligned sentence pairs (Sahlgren & Karlgren 2005)
- No feature weighting or normalisation

➥ **F** is binary, i.e. $f_{ij} \in \{0, 1\}$

Excursus: Similarity in term-context DSM

- What is the cosine similarity of $\mathbf{f}_i$ and $\mathbf{f}_j$?

$$\mathbf{f}_i = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$
$$\mathbf{f}_j = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Excursus: Similarity in term-context DSM

- What is the cosine similarity of $\mathbf{f}_i$ and $\mathbf{f}_j$?

$$\mathbf{f}_i = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$
$$\mathbf{f}_j = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

  - $\mathbf{f}_i^T \mathbf{f}_j = O =$ co-occurrence frequency

Excursus: Similarity in term-context DSM

- What is the cosine similarity of $\mathbf{f}_i$ and $\mathbf{f}_j$?

$$\mathbf{f}_i = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$
$$\mathbf{f}_j = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

  - $\mathbf{f}_i^T \mathbf{f}_j = O =$ co-occurrence frequency
  - $\|\mathbf{f}_i\|_2 = \sqrt{R} =$ marginal frequency of term i

Excursus: Similarity in term-context DSM

- What is the cosine similarity of $\mathbf{f}_i$ and $\mathbf{f}_j$?

$$\mathbf{f}_i = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$
$$\mathbf{f}_j = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

  - $\mathbf{f}_i^T \mathbf{f}_j = O =$ co-occurrence frequency
  - $\|\mathbf{f}_i\|_2 = \sqrt{R} =$ marginal frequency of term i
  - $\|\mathbf{f}_j\|_2 = \sqrt{C} =$ marginal frequency of term j

Excursus: Similarity in term-context DSM

- What is the cosine similarity of $\mathbf{f}_i$ and $\mathbf{f}_j$?

$$\mathbf{f}_i = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$
$$\mathbf{f}_j = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

  - $\mathbf{f}_i^T \mathbf{f}_j = O =$ co-occurrence frequency
  - $\|\mathbf{f}_i\|_2 = \sqrt{R} =$ marginal frequency of term i
  - $\|\mathbf{f}_j\|_2 = \sqrt{C} =$ marginal frequency of term j

- Cosine similarity in $\mathbf{F}$ = **first-order association**

$$\cos\alpha = \frac{\mathbf{f}_i^T \mathbf{f}_j}{\|\mathbf{f}_i\|_2 \cdot \|\mathbf{f}_j\|_2} = \frac{O}{\sqrt{RC}} \sim \sqrt{\mathrm{MI}^2}$$

Excursus: Term-context vs. term-term DSM

- Construct a term-term DSM with textual context = tweet
- Recall: co-occurrence frequency $m_{ij} = \mathbf{f}_i^T \mathbf{f}_j$

Excursus: Term-context vs. term-term DSM

- Construct a term-term DSM with textual context = tweet
- Recall: co-occurrence frequency $m_{ij} = \mathbf{f}_i^T \mathbf{f}_j$
- Symmetric co-occurrence matrix **M** can be derived from **F**:

$$\mathbf{M} = \mathbf{F}\mathbf{F}^T$$

Excursus: Term-context vs. term-term DSM

- Construct a term-term DSM with textual context = tweet
- Recall: co-occurrence frequency $m_{ij} = \mathbf{f}_i^T \mathbf{f}_j$
- Symmetric co-occurrence matrix $\mathbf{M}$ can be derived from $\mathbf{F}$:

$$\mathbf{M} = \mathbf{F}\mathbf{F}^T$$

- Compare SVD of the two matrices

$$\mathbf{F} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \qquad \begin{aligned} \mathbf{M} &= \mathbf{F}\mathbf{F}^T = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T \\ &= \mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^T \end{aligned}$$

Excursus: Term-context vs. term-term DSM

- Construct a term-term DSM with textual context = tweet
- Recall: co-occurrence frequency $m_{ij} = \mathbf{f}_i^T \mathbf{f}_j$
- Symmetric co-occurrence matrix $\mathbf{M}$ can be derived from $\mathbf{F}$:

$$\mathbf{M} = \mathbf{F}\mathbf{F}^T$$

- Compare SVD of the two matrices

$$\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \qquad \begin{aligned} \mathbf{M} &= \mathbf{F}\mathbf{F}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T \\ &= \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T \end{aligned}$$

➥ $\mathbf{M}$ is power-scaled version of $\mathbf{F}$
  - dimensionality reduction: $P_r(\mathbf{F}) = \mathbf{U}_r\mathbf{\Sigma}_r$ vs. $P_r(\mathbf{M}) = \mathbf{U}_r\mathbf{\Sigma}_r^2$
  - $\mathbf{F}$ is equivalent to $\mathbf{M}$ with Caron $P = \frac{1}{2}$

References I

Boutsidis, C. and Gallopoulos, E. (2008). SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, **41**, 1350–1362.

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.

Clark, H.H. (1970). Word associations and linguistic theory. In J. Lyons (ed.), *New horizons in linguistics*. Harmondsworth: Penguin.

Curran, James; Clark, Stephen; Bos, Johan (2007). Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, pages 33–36, Prague, Czech Republic.

Deisenroth, Marc Peter; Faisal, A. Aldo; Ong, Cheng Soon (2020). *Mathematics for Machine Learning*. Cambridge University Press. `https://mml-book.github.io/`.

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

References II

Evert, Stefan and Lapesa, Gabriella (2021). FAST: A carefully sampled and cognitively motivated dataset for distributional semantic evaluation. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL 2021)*, pages 588–595, Online. Data set: `https://osf.io/cd8ar/`.

Ghosh, Urmi; Jain, Sambhav; Paul, Soma (2015). A two-stage approach for computing associative responses to a set of stimulus words. In Z. (eds.) (ed.), *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon,*.

Hare, Mary; Jones, Michael; Thomson, Caroline; Kelly, Sarah; McRae, Ken (2009). Activating event knowledge. *Cognition*, **111**(2), 151–167.

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.

James, W (1890). *The principles of psychology*. New York: Dover.

Joulin, Armand; Grave, Edouard; Bojanowski, Piotr; Mikolov, Tomas (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain.

References III

Kiss, G.R; Armstrong, C.; Milroy; Piper, J. (1973). An associative thesaurus of english and its computer analysis. In R. B. Aitken and N. Hamilton-Smith (eds.), *The computer and literary studies*. Edinburgh University Pres.

Lapesa, Gabriella and Evert, Stefan (2013). Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2013)*, pages 66–74, Sofia, Bulgaria.

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

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

Lee, Daniel D. and Seung, H. Sebastian (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13: Proceedings of the NIPS 2000 Conference*, pages 556–562. MIT Press.

References IV

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

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

Murphy, Brian; Baroni, Marco; Poesio, Massimo (2009). EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 619–627, Singapore.

Nelson, Douglas L.; McEvoy, Cathy L.; Schreiber, Thomas A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*.

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

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

References V

Rapp, Reinhard and Zock, Michael (2014). The cogalex-iv shared task on the lexical access problem. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon,*, pages 1–14. Zock/Rapp/Huang (eds.).

Sahlgren, Magnus and Karlgren, Jussi (2005). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, **11**, 327–341.

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.

Schäfer, Roland and Bildhauer, Felix (2012). Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*, pages 486–493, Istanbul, Turkey. ELRA.

Wettler, Manfred; Rapp, Reinhard; Sedlmeier, Peter (2005). Free word associations correspond to contiguities between words in texts*. *Journal of Quantitative Linguistics*, **12**(2–3), 111–122.