Distributional Semantics

Magnus Sahlgren

Pavia, 13 September 2012

Recap

Distributional Semantics and The Distributional Hypothesis

Syntagmatic and paradigmatic similarities

Co-occurrence matrix and distributional vectors

Vector similarity and nearest neighbors

Recap

Distributional Semantics and The Distributional Hypothesis

Syntagmatic and paradigmatic similarities

Co-occurrence matrix and distributional vectors

Vector similarity and nearest neighbors

Recap

Distributional Semantics and The Distributional Hypothesis

Syntagmatic and paradigmatic similarities

Co-occurrence matrix and distributional vectors

Vector similarity and nearest neighbors

Recap

Distributional Semantics and The Distributional Hypothesis

Syntagmatic and paradigmatic similarities

Co-occurrence matrix and distributional vectors

Vector similarity and nearest neighbors

Recap

Words-by-regions matrices and LSA

Words-by-words-matrices and HAL

Dependency-based models

Random Indexing and Random Permutations

Recap

Words-by-regions matrices and LSA

Words-by-words-matrices and HAL

Dependency-based models

Random Indexing and Random Permutations

Recap

Words-by-regions matrices and LSA

Words-by-words-matrices and HAL

Dependency-based models

Random Indexing and Random Permutations

Recap

Words-by-regions matrices and LSA

Words-by-words-matrices and HAL

Dependency-based models

Random Indexing and Random Permutations

Recap

Words-by-regions matrices and LSA

Words-by-words-matrices and HAL

Dependency-based models

Random Indexing and Random Permutations

Today

Evaluation

Applications

Challenges

Today

Evaluation

Applications

Challenges

Today

Evaluation

Applications

Challenges

Today

Evaluation

Applications

Challenges

Evaluation

- Vocabulary tests
- Similarity ratings
- Association norms
- Behavioral data
- BLESS

Evaluation

Vocabulary tests

Similarity ratings

Association norms

Behavioral data

BLESS

Evaluation

Vocabulary tests

Similarity ratings

Association norms

Behavioral data

BLESS

Evaluation

Vocabulary tests

Similarity ratings

Association norms

Behavioral data

BLESS

Evaluation

Vocabulary tests

Similarity ratings

Association norms

Behavioral data

BLESS

Evaluation

Vocabulary tests

Similarity ratings

Association norms

Behavioral data

BLESS

Evaluation

Vocabulary tests

Multiple choice synonym tests (TOEFL, ESL)

target	**alternatives**	**correct**
flawed	lustrous	
	imperfect	$\surd$
	crude	
	tiny	

Evaluation

Vocabulary tests

Multiple choice synonym tests (TOEFL, ESL)

target	**alternatives**	**correct**
flawed	lustrous	
	imperfect	$\surd$
	crude	
	tiny	

Evaluation

Vocabulary tests

Multiple choice synonym tests (TOEFL, ESL)

target	**alternatives**	**correct**
flawed	lustrous	
	imperfect	$\surd$
	crude	
	tiny	

Evaluation

Vocabulary tests

State-of-the-art: 92.50% correct answers (Rapp, 2003)

- Balanced corpus (BNC)
- Stop word filtering and lemmatization
- 2+2 sized context window
- Entropy-based weighting
- Exclude low-frequent words
- Dimension reduction (SVD to 300 dimensions)

Evaluation

Vocabulary tests

State-of-the-art: 92.50% correct answers (Rapp, 2003)

- Balanced corpus (BNC)
- Stop word filtering and lemmatization
- 2+2 sized context window
- Entropy-based weighting
- Exclude low-frequent words
- Dimension reduction (SVD to 300 dimensions)

Evaluation

Vocabulary tests

State-of-the-art: 92.50% correct answers (Rapp, 2003)

- Balanced corpus (BNC)
- Stop word filtering and lemmatization
- 2+2 sized context window
- Entropy-based weighting
- Exclude low-frequent words
- Dimension reduction (SVD to 300 dimensions)

Evaluation

Vocabulary tests

State-of-the-art: 92.50% correct answers (Rapp, 2003)

- Balanced corpus (BNC)
- Stop word filtering and lemmatization
- 2+2 sized context window
- Entropy-based weighting
- Exclude low-frequent words
- Dimension reduction (SVD to 300 dimensions)

Evaluation

Vocabulary tests

State-of-the-art: 92.50% correct answers (Rapp, 2003)

- Balanced corpus (BNC)
- Stop word filtering and lemmatization
- 2+2 sized context window
- Entropy-based weighting
- Exclude low-frequent words
- Dimension reduction (SVD to 300 dimensions)

Evaluation

Vocabulary tests

State-of-the-art: 92.50% correct answers (Rapp, 2003)

- Balanced corpus (BNC)
- Stop word filtering and lemmatization
- 2+2 sized context window
- Entropy-based weighting
- Exclude low-frequent words
- Dimension reduction (SVD to 300 dimensions)

Evaluation

Vocabulary tests

State-of-the-art: 92.50% correct answers (Rapp, 2003)

- Balanced corpus (BNC)
- Stop word filtering and lemmatization
- 2+2 sized context window
- Entropy-based weighting
- Exclude low-frequent words
- Dimension reduction (SVD to 300 dimensions)

Evaluation

Vocabulary tests

State-of-the-art: 92.50% correct answers (Rapp, 2003)

- Balanced corpus (BNC)
- Stop word filtering and lemmatization
- 2+2 sized context window
- Entropy-based weighting
- Exclude low-frequent words
- Dimension reduction (SVD to 300 dimensions)

Evaluation

Thesaurus comparison

Compare the overlap between a word space and a thesaurus (e.g. Roget's)

Compute the correlation between distributional similarity and graph distance in WordNet

(Who is evaluating who?)

Evaluation

Thesaurus comparison

Compare the overlap between a word space and a thesaurus (e.g. Roget's)

Compute the correlation between distributional similarity and graph distance in WordNet

(Who is evaluating who?)

Evaluation

Thesaurus comparison

Compare the overlap between a word space and a thesaurus (e.g. Roget's)

Compute the correlation between distributional similarity and graph distance in WordNet

(Who is evaluating who?)

Evaluation

Thesaurus comparison

Compare the overlap between a word space and a thesaurus (e.g. Roget's)

Compute the correlation between distributional similarity and graph distance in WordNet

(Who is evaluating who?)

Evaluation

Similarity ratings

Rate the similarity betwen given word pairs

word 1	word 2	mean rating
tiger	cat	7.35
sex	love	6.77
stock	cd	1.31

What kind of similarity?

Evaluation

Similarity ratings

Rate the similarity betwen given word pairs

word 1	word 2	mean rating
tiger	cat	7.35
sex	love	6.77
stock	cd	1.31

What kind of similarity?

Evaluation

Similarity ratings

Rate the similarity betwen given word pairs

word 1	**word 2**	**mean rating**
tiger	cat	7.35
sex	love	6.77
stock	cd	1.31

What kind of similarity?

Evaluation

Similarity ratings

Rate the similarity betwen given word pairs

word 1	**word 2**	**mean rating**
tiger	cat	7.35
sex	love	6.77
stock	cd	1.31

What kind of similarity?

Evaluation

Association norms

Name the first word that comes to mind when having seen a stimuli

stimuli	respons	association strength
galaxy	stars	0.413
galaxy	space	0.107
galaxy	far	0.020

Associations tend to be syntagmatic in nature

Evaluation

Association norms

Name the first word that comes to mind when having seen a stimuli

stimuli	respons	association strength
galaxy	stars	0.413
galaxy	space	0.107
galaxy	far	0.020

Associations tend to be syntagmatic in nature

Evaluation

Association norms

Name the first word that comes to mind when having seen a stimuli

stimuli	**respons**	**association strength**
galaxy	stars	0.413
galaxy	space	0.107
galaxy	far	0.020

Associations tend to be syntagmatic in nature

Evaluation

Association norms

Name the first word that comes to mind when having seen a stimuli

stimuli	**respons**	**association strength**
galaxy	stars	0.413
galaxy	space	0.107
galaxy	far	0.020

Associations tend to be syntagmatic in nature

Evaluation

Semantic Priming

Hearing/reading a "related" prime facilitates access to a target in various lexical tasks (naming, lexical decision, reading)

The word *pear* is recognized/accessed faster if it is heard/read after *apple*

Evaluation

Semantic Priming

Hearing/reading a “related” prime facilitates access to a target in various lexical tasks (naming, lexical decision, reading)

The word *pear* is recognized/accessed faster if it is heard/read after *apple*

Evaluation
Semantic Priming

Modelling semantic priming in word space

1. Measure the similarity between the distributional vectors of each prime-target pair
2. The similarity between related items should be significantly higher than similarity between unrelated items

Evaluation

Semantic Priming

Modelling semantic priming in word space

1. Measure the similarity between the distributional vectors of each prime-target pair
2. The similarity between related items should be significantly higher than similarity between unrelated items

Evaluation

Semantic Priming

Modelling semantic priming in word space

1. Measure the similarity between the distributional vectors of each prime-target pair
2. The similarity between related items should be significantly higher than similarity between unrelated items

Evaluation

Semantic Priming

Hodgson (1991) single word lexical decision task, 136 prime-target pairs

Similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation):

- synonyms: *to dread/to fear*
- antonyms: *short/tall*
- coordinates: *train/truck*
- super- and subordinate pairs: *container/bottle*
- free association pairs: *dove/peace*
- phrasal associates: *vacant/building*

Evaluation

Semantic Priming

Hodgson (1991) single word lexical decision task, 136 prime-target pairs

Similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation):

- synonyms: *to dread/to fear*
- antonyms: *short/tall*
- coordinates: *train/truck*
- super- and subordinate pairs: *container/bottle*
- free association pairs: *dove/peace*
- phrasal associates: *vacant/building*

Evaluation

Semantic Priming

Hodgson (1991) single word lexical decision task, 136 prime-target pairs

Similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation):

- synonyms: *to dread/to fear*
- antonyms: *short/tall*
- coordinates: *train/truck*
- super- and subordinate pairs: *container/bottle*
- free association pairs: *dove/peace*
- phrasal associates: *vacant/building*

Evaluation

Semantic Priming

Hodgson (1991) single word lexical decision task, 136 prime-target pairs

Similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation):

- synonyms: *to dread/to fear*
- antonyms: *short/tall*
- coordinates: *train/truck*
- super- and subordinate pairs: *container/bottle*
- free association pairs: *dove/peace*
- phrasal associates: *vacant/building*

Evaluation

Semantic Priming

Hodgson (1991) single word lexical decision task, 136 prime-target pairs

Similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation):

- synonyms: *to dread/to fear*
- antonyms: *short/tall*
- coordinates: *train/truck*
- super- and subordinate pairs: *container/bottle*
- free association pairs: *dove/peace*
- phrasal associates: *vacant/building*

Evaluation

Semantic Priming

Hodgson (1991) single word lexical decision task, 136 prime-target pairs

Similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation):

- synonyms: *to dread/to fear*
- antonyms: *short/tall*
- coordinates: *train/truck*
- super- and subordinate pairs: *container/bottle*
- free association pairs: *dove/peace*
- phrasal associates: *vacant/building*

Evaluation

Semantic Priming

Hodgson (1991) single word lexical decision task, 136 prime-target pairs

Similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation):

- synonyms: *to dread/to fear*
- antonyms: *short/tall*
- coordinates: *train/truck*
- super- and subordinate pairs: *container/bottle*
- free association pairs: *dove/peace*
- phrasal associates: *vacant/building*

Evaluation

Semantic Priming

Hodgson (1991) single word lexical decision task, 136 prime-target pairs

Similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation):

- synonyms: *to dread/to fear*
- antonyms: *short/tall*
- coordinates: *train/truck*
- super- and subordinate pairs: *container/bottle*
- free association pairs: *dove/peace*
- phrasal associates: *vacant/building*

Evaluation

Semantic Priming

Dependency-based model (Padó & Lapata 2007)

Mean distance values for Related and Unrelated prime–target pairs; Prime Effect size (= Related − Unrelated) for the dependency model and ICE.

Lexical Relation	N	Related	Unrelated	Effect (dependency)	Effect (ICE)
Synonymy	23	0.267	0.102	0.165**	0.063
Superordination	21	0.227	0.121	0.106**	0.067
Category coordination	23	0.256	0.119	0.137**	0.074
Antonymy	24	0.292	0.127	0.165**	0.097
Conceptual association	23	0.204	0.121	0.083**	0.086
Phrasal association	22	0.146	0.103	0.043**	0.058

**$p < 0.01$ (2-tailed)

Evaluation

BLESS (Baroni & Lenci, 2011)

26 554 tuples expressing a *relation* between a target concept and a relatum

200 basic-level nominal concrete concepts, 8 relation types, each instantiated by multiple relata (nouns, verbs or adjectives)

Evaluation

BLESS (Baroni & Lenci, 2011)

26 554 tuples expressing a *relation* between a target concept and a relatum

200 basic-level nominal concrete concepts, 8 relation types, each instantiated by multiple relata (nouns, verbs or adjectives)

Evaluation

BLESS (Baroni & Lenci, 2011)

26 554 tuples expressing a *relation* between a target concept and a relatum

200 basic-level nominal concrete concepts, 8 relation types, each instantiated by multiple relata (nouns, verbs or adjectives)

Evaluation

BLESS (Baroni & Lenci, 2011)

Target concepts are 200 English concrete nouns (100 living and 100 non-living) grouped into 17 broader classes

`amphibian_reptile, appliance, bird, building, clothing, container, fruit, furniture, ground_mammal, insect, musical_instrument, tool, tree, vehicle, water_animal, weapon`

Evaluation

BLESS (Baroni & Lenci, 2011)

Target concepts are 200 English concrete nouns (100 living and 100 non-living) grouped into 17 broader classes

`amphibian_reptile, appliance, bird, building, clothing, container, fruit, furniture, ground_mammal, insect, musical_instrument, tool, tree, vehicle, water_animal, weapon`

Evaluation

BLESS (Baroni & Lenci, 2011)

Relations

- **coord** relatum is a co-hyponym (coordinate) of the target (`guitar, coord, violin`)
- **hyper** relatum is a hypernym of the target (`rabbit, hyper, animal`)
- **mero** relatum is a noun referring to a part of the target (`beaver, mero, fur`)
- **attri** relatum expresses an attribute of the target (`sword, attri, dangerous`)
- **event** relatum expresses an event involving the target (`butterfly, event, fly`)
- **ran.k** relatum is a random noun ($k = n$), adjective ($k = j$) and verb ($k = v$) (`donkey, ran.`v`, coincide`)

Evaluation

BLESS (Baroni & Lenci, 2011)

Relations

- **coord** relatum is a co-hyponym (coordinate) of the target (`guitar, coord, violin`)
- **hyper** relatum is a hypernym of the target (`rabbit, hyper, animal`)
- **mero** relatum is a noun referring to a part of the target (`beaver, mero, fur`)
- **attri** relatum expresses an attribute of the target (`sword, attri, dangerous`)
- **event** relatum expresses an event involving the target (`butterfly, event, fly`)
- **ran.k** relatum is a random noun ($k = n$), adjective ($k = j$) and verb ($k = v$) (donkey, ran.v, coincide)

Evaluation

BLESS (Baroni & Lenci, 2011)

target	**relation**	**relata**
rabbit	hyper	*animal, chordate, mammal, ...*
guitar	coord	*violin, trumpet, piano, ...*
beaver	mero	*fur, head, tooth, ...*
sword	attri	*dangerous, long, heavy, ...*
butterfly	event	*fly, catch, flutter, ...*
villa	ran.n	*disease, assistance, game, ...*
donkey	ran.v	*coincide, express, vent, ...*
hat	ran.j	*quarterly, massive, obvious, ...*

Evaluation

BLESS (Baroni & Lenci, 2011)

For each of the 200 BLESS concepts, 8 similarity scores are computed, by picking the relatum with the highest similarity for each relation

The 8 similarity scores are transformed onto standardized z-scores to account for frequency effects

For each relation type, the distribution of scores across the 200 concepts is summarized with a *boxplot*

Evaluation

BLESS (Baroni & Lenci, 2011)

For each of the 200 BLESS concepts, 8 similarity scores are computed, by picking the relatum with the highest similarity for each relation

The 8 similarity scores are transformed onto standardized z-scores to account for frequency effects

For each relation type, the distribution of scores across the 200 concepts is summarized with a *boxplot*

Evaluation

BLESS (Baroni & Lenci, 2011)

For each of the 200 BLESS concepts, 8 similarity scores are computed, by picking the relatum with the highest similarity for each relation

The 8 similarity scores are transformed onto standardized z-scores to account for frequency effects

For each relation type, the distribution of scores across the 200 concepts is summarized with a *boxplot*

Evaluation

BLESS (Baroni & Lenci, 2011)

For each of the 200 BLESS concepts, 8 similarity scores are computed, by picking the relatum with the highest similarity for each relation

The 8 similarity scores are transformed onto standardized *z*-scores to account for frequency effects

For each relation type, the distribution of scores across the 200 concepts is summarized with a *boxplot*

Evaluation

BLESS (Baroni & Lenci, 2011)

Document: a words-by-documents matrix

ContentWindow20: a 20+20-sized context window

ContentWindow2: a 2+2-sized context window

Evaluation

BLESS (Baroni & Lenci, 2011)

Document: a words-by-documents matrix

ContentWindow20: a 20+20-sized context window

ContentWindow2: a 2+2-sized context window

Evaluation

BLESS (Baroni & Lenci, 2011)

Document: a words-by-documents matrix

ContentWindow20: a 20+20-sized context window

ContentWindow2: a 2+2-sized context window

Evaluation

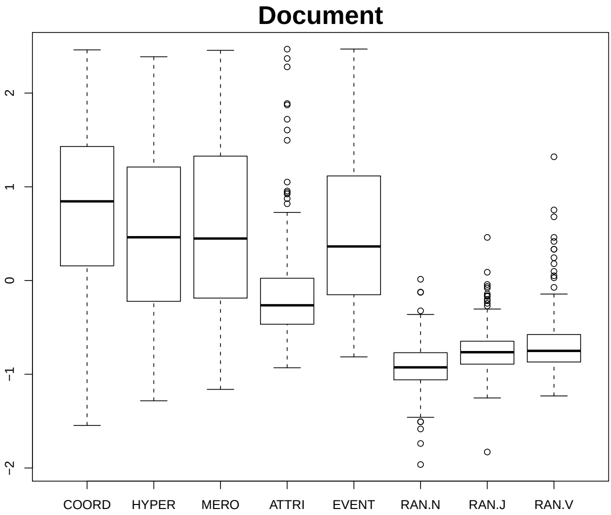
BLESS (Baroni & Lenci, 2011)

Document: a words-by-documents matrix

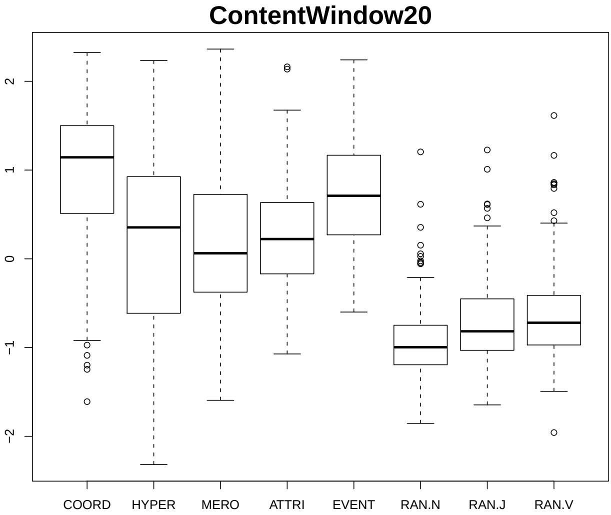
ContentWindow20: a 20+20-sized context window

ContentWindow2: a 2+2-sized context window

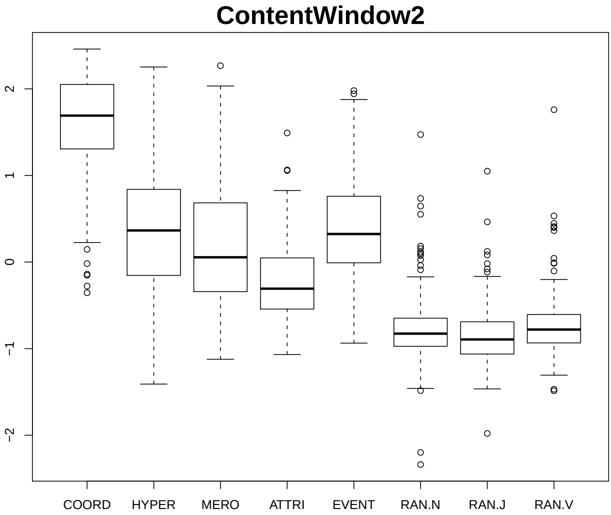
Evaluation



Evaluation


ContentWindow20
2
1
0
−1
−2
COORD
HYPER
MERO
ATTRI
EVENT
RAN.N
RAN.J
RAN.V

Evaluation


ContentWindow2
2
1
0
−1
−2
COORD
HYPER
MERO
ATTRI
EVENT
RAN.N
RAN.J
RAN.V

Evaluation

Whatever we find in the data is the truth (about that data)

Applications

- Automatically acquiring (multilingual) lexical resources
- Word sense discrimination
- Selectional preferences
- ...

Applications

- Automatically acquiring (multilingual) lexical resources
- Word sense discrimination
- Selectional preferences
- ...

Applications

- Automatically acquiring (multilingual) lexical resources
- Word sense discrimination
- Selectional preferences
- ...

Applications

- Automatically acquiring (multilingual) lexical resources
- Word sense discrimination
- Selectional preferences
- ...

Applications

- Automatically acquiring (multilingual) lexical resources
- Word sense discrimination
- Selectional preferences
- ...

Multilingual Word Spaces

Different word spaces with aligned contexts

(Can also be used across domains and genres)

Use parallel data and train a words-by-documents model

Multilingual Word Spaces

Different word spaces with aligned contexts

(Can also be used across domains and genres)

Use parallel data and train a words-by-documents model

Multilingual Word Spaces

Different word spaces with aligned contexts

(Can also be used across domains and genres)

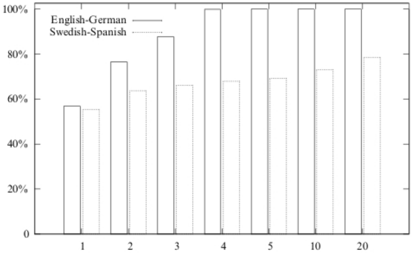
Use parallel data and train a words-by-documents model

Multilingual Word Spaces

SWEDISH	ITALIAN
Jag förklarar Europaparlamentets session återupptagen efter avbrottet den 17 december . Jag vill på nytt önska er ett gott nytt år och jag hoppas att ni haft en trevlig semester .	Dichiaro ripresa la sessione del Parlamento europeo , interrotta venerdì 17 dicembre e rinnovo a tutti i miei migliori auguri nella speranza che abbiate trascorso delle buone vacanze .
Som ni kunnat konstatera ägde den stora år 2000-buggen aldrig rum .	Come avrete avuto modo di constatare il grande baco del millennio non si è materializzato .

Multilingual Word Spaces

Single word translations with different number of alternatives (Sahlgren & Karlgren, 2005)



Word sense discrimination

Schütze 1998

Build different vectors for different senses

Context vector:

- for each word token w_i, take the words in its context C_i

  $C_1 = \{\texttt{cat, chase}\}$
  $C_2 = \{\texttt{hacker, click, button}\}$

- for each C_i, build a context vector $\overrightarrow{C_i}$ by summing thedistributional vectors of the words in C_i

  $\overrightarrow{C_1} = \overrightarrow{\texttt{cat}} + \overrightarrow{\texttt{chase}}$
  $\overrightarrow{C_2} = \overrightarrow{\texttt{hacker}} + \overrightarrow{\texttt{click}} + \overrightarrow{\texttt{button}}$

Word sense discrimination

Schütze 1998

Build different vectors for different senses

Context vector:

- for each word token w_i, take the words in its context C_i

  $C_1 = \{\texttt{cat}, \texttt{chase}\}$
  $C_2 = \{\texttt{hacker}, \texttt{click}, \texttt{button}\}$

- for each C_i, build a context vector $\overrightarrow{C_i}$ by summing thedistributional vectors of the words in C_i

  $\overrightarrow{C_1} = \overrightarrow{\texttt{cat}} + \overrightarrow{\texttt{chase}}$
  $\overrightarrow{C_2} = \overrightarrow{\texttt{hacker}} + \overrightarrow{\texttt{click}} + \overrightarrow{\texttt{button}}$

Word sense discrimination

Schütze 1998

Build different vectors for different senses

Context vector:

- for each word token w_i, take the words in its context C_i

  $C_1 = \{\texttt{cat}, \texttt{chase}\}$

  $C_2 = \{\texttt{hacker}, \texttt{click}, \texttt{button}\}$

- for each C_i, build a context vector $\overrightarrow{C_i}$ by summing thedistributional vectors of the words in C_i

  $\overrightarrow{C_1} = \overrightarrow{\texttt{cat}} + \overrightarrow{\texttt{chase}}$

  $\overrightarrow{C_2} = \overrightarrow{\texttt{hacker}} + \overrightarrow{\texttt{click}} + \overrightarrow{\texttt{button}}$

Word sense discrimination

Schütze 1998

Build different vectors for different senses

Context vector:

- for each word token w_i, take the words in its context C_i

  $C_1 = \{\texttt{cat}, \texttt{chase}\}$
  $C_2 = \{\texttt{hacker}, \texttt{click}, \texttt{button}\}$

- for each C_i, build a context vector $\overrightarrow{C_i}$ by summing thedistributional vectors of the words in C_i

  $\overrightarrow{C_1} = \overrightarrow{\texttt{cat}} + \overrightarrow{\texttt{chase}}$
  $\overrightarrow{C_2} = \overrightarrow{\texttt{hacker}} + \overrightarrow{\texttt{click}} + \overrightarrow{\texttt{button}}$

Word sense discrimination

Schütze 1998

Build different vectors for different senses

Context vector:

- for each word token w_i, take the words in its context C_i

  $C_1 = \{\texttt{cat}, \ \texttt{chase}\}$
  $C_2 = \{\texttt{hacker}, \ \texttt{click}, \ \texttt{button}\}$

- for each C_i, build a context vector $\overrightarrow{C_i}$ by summing thedistributional vectors of the words in C_i

  $\overrightarrow{C_1} = \overrightarrow{\texttt{cat}} + \overrightarrow{\texttt{chase}}$
  $\overrightarrow{C_2} = \overrightarrow{\texttt{hacker}} + \overrightarrow{\texttt{click}} + \overrightarrow{\texttt{button}}$

Word sense discrimination

Schütze 1998

Build different vectors for different senses

Context vector:

- for each word token w_i, take the words in its context C_i

  $$C_1 = \{\texttt{cat}, \ \texttt{chase}\}$$
  $$C_2 = \{\texttt{hacker}, \ \texttt{click}, \ \texttt{button}\}$$

- for each C_i, build a context vector $\overrightarrow{C_i}$ by summing thedistributional vectors of the words in C_i

  $$\overrightarrow{C_1} = \overrightarrow{\texttt{cat}} + \overrightarrow{\texttt{chase}}$$
  $$\overrightarrow{C_2} = \overrightarrow{\texttt{hacker}} + \overrightarrow{\texttt{click}} + \overrightarrow{\texttt{button}}$$

Word sense discrimination

Schütze 1998

Build different vectors for different senses

Context vector:

- for each word token w_i, take the words in its context C_i

  $C_1 = \{\texttt{cat}, \ \texttt{chase}\}$
  $C_2 = \{\texttt{hacker}, \ \texttt{click}, \ \texttt{button}\}$

- for each C_i, build a context vector $\overrightarrow{C_i}$ by summing thedistributional vectors of the words in C_i

  $\overrightarrow{C_1} = \overrightarrow{\texttt{cat}} + \overrightarrow{\texttt{chase}}$
  $\overrightarrow{C_2} = \overrightarrow{\texttt{hacker}} + \overrightarrow{\texttt{click}} + \overrightarrow{\texttt{button}}$

Word sense discrimination

Schütze 1998

A context vector is the *centroid* of the distributional vectors of the context words

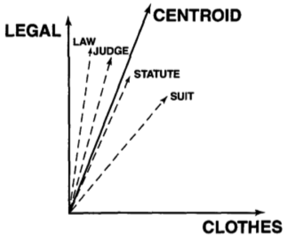


Figure 3
The derivation of context vectors. A context vector is computed as the centroid of the words occurring in the context. The words in this example context are *law*, *judge*, *statute*, and *suit*.

Word sense discrimination

Schütze 1998

Word senses are represented by clusters of similar contexts

1. take all the contexts of a word w in a training corpus
2. build the context vector $\overrightarrow{C_i}$, for each of these contexts
3. cluster the context vectors
4. for each cluster, take the *centroid vector* of the cluster, and use this vector to represent one *sense* of w (*sense vector*, $\overrightarrow{s_j}$)

Word sense discrimination

Schütze 1998

Word senses are represented by clusters of similar contexts

1. take all the contexts of a word w in a training corpus
2. build the context vector $\overrightarrow{C_i}$, for each of these contexts
3. cluster the context vectors
4. for each cluster, take the *centroid vector* of the cluster, and use this vector to represent one *sense* of w (*sense vector*, $\overrightarrow{s_j}$)

Word sense discrimination

Schütze 1998

Word senses are represented by clusters of similar contexts

1. take all the contexts of a word w in a training corpus
2. build the context vector $\overrightarrow{C_i}$, for each of these contexts
3. cluster the context vectors
4. for each cluster, take the *centroid vector* of the cluster, and use this vector to represent one *sense* of w (*sense vector*, $\overrightarrow{s_j}$)

Word sense discrimination

Schütze 1998

Word senses are represented by clusters of similar contexts

1. take all the contexts of a word w in a training corpus
2. build the context vector $\overrightarrow{C_i}$, for each of these contexts
3. cluster the context vectors
4. for each cluster, take the *centroid vector* of the cluster, and use this vector to represent one *sense* of w (*sense vector*, $\overrightarrow{s_j}$)

Word sense discrimination

Schütze 1998

Word senses are represented by clusters of similar contexts

1. take all the contexts of a word w in a training corpus
2. build the context vector $\overrightarrow{C_i}$, for each of these contexts
3. cluster the context vectors
4. for each cluster, take the *centroid vector* of the cluster, and use this vector to represent one *sense* of w (*sense vector*, $\overrightarrow{s_j}$)

Word sense discrimination

Schütze 1998

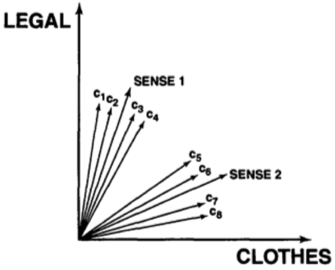


Figure 4
The derivation of sense vectors. Sense vectors are derived by clustering the context vectors of an ambiguous word (here, $c_1, c_2, c_3, c_4, c_5, c_6, c_7$, and c_8), and computing sense vectors as the centroids of the resulting clusters. The vectors SENSE 1 and SENSE 2 are the sense vectors of clusters $\{c_1, c_2, c_3, c_4\}$ and $\{c_5, c_6, c_7, c_8\}$, respectively.

Word sense discrimination

Schütze 1998

To assign a sense to a new instance of w in context C_k

1. build the context vector $\overrightarrow{C_k}$
2. assign to w in context C_k the sense j whose sense vector $\overrightarrow{s_j}$ is closest to $\overrightarrow{C_k}$

Word sense discrimination

Schütze 1998

To assign a sense to a new instance of w in context C_k

1. build the context vector $\overrightarrow{C_k}$
2. assign to w in context C_k the sense j whose sense vector $\overrightarrow{s_j}$ is closest to $\overrightarrow{C_k}$

Word sense discrimination

Schütze 1998

To assign a sense to a new instance of w in context C_k

1. build the context vector $\overrightarrow{C_k}$
2. assign to w in context C_k the sense j whose sense vector $\overrightarrow{s_j}$ is closest to $\overrightarrow{C_k}$

Word sense discrimination

Schütze 1998

Level	Average of $2 \times sin^{-1}(\sqrt{X})$	Difference from Closest	Corresponding Accuracy
local, χ^2, terms	2.11	0.13	76%
local, frequency, terms	2.24	0.13	81%
local, frequency, SVD	2.44	0.06	88%
local, χ^2, SVD	2.50	0.06	90%
global, frequency, SVD	2.66	0.16	94%

Selectional Preferences

Selectional preferences specify an abstract semantic type constraining the possible arguments of a predicate

- `kill-obj`: living_entity
- `eat-obj`: food
- `drink-obj`: liquid

Necessary to account for the possibility of generalizations to unseen arguments

Selectional Preferences

Selectional preferences specify an abstract semantic type constraining the possible arguments of a predicate

- `kill-obj`: living_entity
- `eat-obj`: food
- `drink-obj`: liquid

Necessary to account for the possibility of generalizations to unseen arguments

Selectional Preferences

Selectional preferences specify an abstract semantic type constraining the possible arguments of a predicate

- `kill`-obj: living_entity
- `eat`-obj: food
- `drink`-obj: liquid

Necessary to account for the possibility of generalizations to unseen arguments

Selectional Preferences

Selectional preferences specify an abstract semantic type constraining the possible arguments of a predicate

- `kill`-obj: living_entity
- eat-obj: food
- `drink`-obj: liquid

Necessary to account for the possibility of generalizations to unseen arguments

Selectional Preferences

Selectional preferences specify an abstract semantic type constraining the possible arguments of a predicate

- `kill`-obj: living_entity
- `eat`-obj: food
- `drink`-obj: liquid

Necessary to account for the possibility of generalizations to unseen arguments

Selectional Preferences

Selectional preferences specify an abstract semantic type constraining the possible arguments of a predicate

- `kill-obj`: living_entity
- `eat-obj`: food
- `drink-obj`: liquid

Necessary to account for the possibility of generalizations to unseen arguments

Selectional Preferences

We can determine the plausibility of the following phrases, despite never have encountered them in any corpus:

`eat the aardvark`
(`aardvark` is a living entity)

`eat the serendipity`
(`serendipity` is not a living entity)

Selectional Preferences

We can determine the plausibility of the following phrases, despite never have encountered them in any corpus:

`eat the aardvark`
(`aardvark` is a living entity)

`eat the serendipity`
(`serendipity` is not a living entity)

Selectional Preferences

We can determine the plausibility of the following phrases, despite never have encountered them in any corpus:

`eat the aardvark`
(`aardvark` is a living entity)

`eat the serendipity`
(`serendipity` is not a living entity)

Selectional Preferences

We can determine the plausibility of the following phrases, despite never have encountered them in any corpus:

`eat the aardvark`
(`aardvark` is a living entity)

`eat the serendipity`
(`serendipity` is not a living entity)

Selectional Preferences

We can determine the plausibility of the following phrases, despite never have encountered them in any corpus:

`eat the aardvark`
(`aardvark` is a living entity)

`eat the serendipity`
(`serendipity` is not a living entity)

Selectional Preferences

Semantic types are normally extracted from some ontology (e.g. WordNet)

Plausibilities are determined based on the semantic types

Selectional Preferences

Semantic types are normally extracted from some ontology (e.g. WordNet)

Plausibilities are determined based on the semantic types

Selectional Preferences

Behavioral evidence (e.g., semantic priming) suggests that verbs and their arguments are arranged into a web of mutual expectations in the mental lexicon

- verbs activate expectations about prototypical noun arguments (Ferretti et al. 2001)
  `eat-v` $\rightarrow$ `cheese-n`
- nouns activate expectations about prototypical verbs they are arguments of (McRae et al. 2005)
  `juice-n` $\rightarrow$ `drink-v`
- nouns activate expectations about other nouns with which they are related by some event (Hare et al. 2009)
  `key-n` $\rightarrow$ `door-n`

Selectional Preferences

Behavioral evidence (e.g., semantic priming) suggests that verbs and their arguments are arranged into a web of mutual expectations in the mental lexicon

- verbs activate expectations about prototypical noun arguments (Ferretti et al. 2001)
  `eat-v` $\rightarrow$ `cheese-n`
- nouns activate expectations about prototypical verbs they are arguments of (McRae et al. 2005)
  `juice-n` $\rightarrow$ `drink-v`
- nouns activate expectations about other nouns with which they are related by some event (Hare et al. 2009)
  `key-n` $\rightarrow$ `door-n`

Selectional Preferences

Behavioral evidence (e.g., semantic priming) suggests that verbs and their arguments are arranged into a web of mutual expectations in the mental lexicon

- verbs activate expectations about prototypical noun arguments (Ferretti et al. 2001)
  `eat-v` $\rightarrow$ `cheese-n`
- nouns activate expectations about prototypical verbs they are arguments of (McRae et al. 2005)
  `juice-n` $\rightarrow$ `drink-v`
- nouns activate expectations about other nouns with which they are related by some event (Hare et al. 2009)
  `key-n` $\rightarrow$ `door-n`

Selectional Preferences

Behavioral evidence (e.g., semantic priming) suggests that verbs and their arguments are arranged into a web of mutual expectations in the mental lexicon

- verbs activate expectations about prototypical noun arguments (Ferretti et al. 2001)
  `eat-v` $\rightarrow$ `cheese-n`
- nouns activate expectations about prototypical verbs they are arguments of (McRae et al. 2005)
  `juice-n` $\rightarrow$ `drink-v`
- nouns activate expectations about other nouns with which they are related by some event (Hare et al. 2009)
  `key-n` $\rightarrow$ `door-n`

Selectional Preferences

Behavioral evidence (e.g., semantic priming) suggests that verbs and their arguments are arranged into a web of mutual expectations in the mental lexicon

- verbs activate expectations about prototypical noun arguments (Ferretti et al. 2001)
  `eat-v` $\rightarrow$ `cheese-n`
- nouns activate expectations about prototypical verbs they are arguments of (McRae et al. 2005)
  `juice-n` $\rightarrow$ `drink-v`
- nouns activate expectations about other nouns with which they are related by some event (Hare et al. 2009)
  `key-n` $\rightarrow$ `door-n`

Selectional Preferences

Behavioral evidence (e.g., semantic priming) suggests that verbs and their arguments are arranged into a web of mutual expectations in the mental lexicon

- verbs activate expectations about prototypical noun arguments (Ferretti et al. 2001)
  `eat-v` $\rightarrow$ `cheese-n`
- nouns activate expectations about prototypical verbs they are arguments of (McRae et al. 2005)
  `juice-n` $\rightarrow$ `drink-v`
- nouns activate expectations about other nouns with which they are related by some event (Hare et al. 2009)
  `key-n` $\rightarrow$ `door-n`

Selectional Preferences

Behavioral evidence (e.g., semantic priming) suggests that verbs and their arguments are arranged into a web of mutual expectations in the mental lexicon

- verbs activate expectations about prototypical noun arguments (Ferretti et al. 2001)
  `eat-v` $\rightarrow$ `cheese-n`
- nouns activate expectations about prototypical verbs they are arguments of (McRae et al. 2005)
  `juice-n` $\rightarrow$ `drink-v`
- nouns activate expectations about other nouns with which they are related by some event (Hare et al. 2009)
  `key-n` $\rightarrow$ `door-n`

Selectional Preferences

Selectional preferences are *gradual*

`arrest a thief`	very probable
`arrest a policeman`	possible, but less probable
`arrest a tree`	improbable

Selectional Preferences

Selectional preferences are *gradual*

`arrest a thief`	very probable
`arrest a policeman`	possible, but less probable
`arrest a tree`	improbable

Selectional Preferences

Two sources of expectations determining thematic fit judgments:

- Physical experience
- Distributional regularities (i.e. co-occurrences between verbs and arguments)

Selectional Preferences

Two sources of expectations determining thematic fit judgments:

- Physical experience
- Distributional regularities (i.e. co-occurrences between verbs and arguments)

Selectional Preferences

Two sources of expectations determining thematic fit judgments:

- Physical experience
- Distributional regularities (i.e. co-occurrences between verbs and arguments)

Selectional Preferences

Distributional modelling of selectional preferences

The thematic fit of a noun n as an argument of a verb v is modelled as the distributional similarity between n and the most *expected* arguments of v

- Exemplar-based (Erk et al. 2010)
- Prototype-based (Baroni & Lenci 2010)

Selectional Preferences

Distributional modelling of selectional preferences

The thematic fit of a noun n as an argument of a verb v is modelled as the distributional similarity between n and the most *expected* arguments of v

- Exemplar-based (Erk et al. 2010)
- Prototype-based (Baroni & Lenci 2010)

Selectional Preferences

Distributional modelling of selectional preferences

The thematic fit of a noun n as an argument of a verb v is modelled as the distributional similarity between n and the most *expected* arguments of v

- Exemplar-based (Erk et al. 2010)
- Prototype-based (Baroni & Lenci 2010)

Selectional Preferences

Distributional modelling of selectional preferences

The thematic fit of a noun n as an argument of a verb v is modelled as the distributional similarity between n and the most *expected* arguments of v

- Exemplar-based (Erk et al. 2010)
- Prototype-based (Baroni & Lenci 2010)

Selectional Preferences

Using a word space to determine the plausibility of a new noun *n*

Compute the similarity between the distributional vector of *n* and:

- All exemplars of the predicate slot
- The prototye vector of the predicate slot

Selectional Preferences

Using a word space to determine the plausibility of a new noun n

Compute the similarity between the distributional vector of n and:

- All exemplars of the predicate slot
- The prototye vector of the predicate slot

Selectional Preferences

Using a word space to determine the plausibility of a new noun n

Compute the similarity between the distributional vector of n and:

- All exemplars of the predicate slot
- The prototye vector of the predicate slot

Selectional Preferences

Using a word space to determine the plausibility of a new noun n

Compute the similarity between the distributional vector of n and:

- All exemplars of the predicate slot
- The prototye vector of the predicate slot

Selectional Preferences

Human plausibility judgments of noun-verb pairs

shoot	*deer*	obj	6.4
shoot	*deer*	subj	1.0

Selectional Preferences

Human plausibility judgments of noun-verb pairs

shoot	*deer*	obj	6.4
shoot	*deer*	subj	1.0

Selectional Preferences

Performance measured with with Spearman ρ correlation coefficient

MODEL	DATA SET 1	DATA SET 2
WordNet	3	24
Word space	21–41	34–60

Selectional Preferences

Performance measured with with Spearman ρ correlation coefficient

MODEL	DATA SET 1	DATA SET 2
WordNet	3	24
Word space	21–41	34–60

Selectional Preferences

Plausibility of potential objects of `kill`

OBJECT	COSINE
kangaroo	0.51
person	0.45
robot	0.15
hate	0.11
flower	0.11
stone	0.05
fun	0.05
book	0.04
conversation	0.03
sympathy	0.01

Selectional Preferences

Plausibility of potential objects of `kill`

OBJECT	COSINE
kangaroo	0.51
person	0.45
robot	0.15
hate	0.11
flower	0.11
stone	0.05
fun	0.05
book	0.04
conversation	0.03
sympathy	0.01

Other applications

- Document processing (search, categorization, clustering)
- Bioinformatics
- Topic/event detection
- Sentiment analysis
- ...

Other applications

- Document processing (search, categorization, clustering)
- Bioinformatics
- Topic/event detection
- Sentiment analysis
- ...

Other applications

- Document processing (search, categorization, clustering)
- Bioinformatics
- Topic/event detection
- Sentiment analysis
- ...

Other applications

- Document processing (search, categorization, clustering)
- Bioinformatics
- Topic/event detection
- Sentiment analysis
- ...

Other applications

- Document processing (search, categorization, clustering)
- Bioinformatics
- Topic/event detection
- Sentiment analysis
- ...

Challenges for Distributional Semantics

Polysemy

Compositionality

Semantic relations

Visualization

Challenges for Distributional Semantics

Polysemy

Compositionality

Semantic relations

Visualization

Challenges for Distributional Semantics

Polysemy

Compositionality

Semantic relations

Visualization

Challenges for Distributional Semantics

Polysemy

Compositionality

Semantic relations

Visualization

Polysemy

Trivially solved (or rather ignored...) in formal semantics

`The cat chases the mouse` $\Rightarrow$ $\texttt{mouse}_1$

`The hacker clicks the mouse button` $\Rightarrow$ $\texttt{mouse}_2$

In word space, each word type has one vector

`The cat chases the mouse` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

`The hacker clicks the mouse button` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

Polysemy

Trivially solved (or rather ignored...) in formal semantics

`The cat chases the mouse` $\Rightarrow$ $\texttt{mouse}_1$

`The hacker clicks the mouse button` $\Rightarrow$ $\texttt{mouse}_2$

In word space, each word type has one vector

`The cat chases the mouse` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

`The hacker clicks the mouse button` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

Polysemy

Trivially solved (or rather ignored. . .) in formal semantics

`The cat chases the mouse` $\Rightarrow$ $\texttt{mouse}_1$

`The hacker clicks the mouse button` $\Rightarrow$ $\texttt{mouse}_2$

In word space, each word type has one vector

`The cat chases the mouse` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

`The hacker clicks the mouse button` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

Polysemy

Trivially solved (or rather ignored...) in formal semantics

`The cat chases the mouse` $\Rightarrow$ $\texttt{mouse}_1$

`The hacker clicks the mouse button` $\Rightarrow$ $\texttt{mouse}_2$

In word space, each word type has one vector

`The cat chases the mouse` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

`The hacker clicks the mouse button` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

Polysemy

Trivially solved (or rather ignored. . .) in formal semantics

`The cat chases the mouse` $\Rightarrow$ $\texttt{mouse}_1$

`The hacker clicks the mouse button` $\Rightarrow$ $\texttt{mouse}_2$

In word space, each word type has one vector

`The cat chases the mouse` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

`The hacker clicks the mouse button` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

Polysemy

Trivially solved (or rather ignored...) in formal semantics

`The cat chases the mouse` $\Rightarrow$ $\texttt{mouse}_1$

`The hacker clicks the mouse button` $\Rightarrow$ $\texttt{mouse}_2$

In word space, each word type has one vector

`The cat chases the mouse` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

`The hacker clicks the mouse button` $\Rightarrow$ $\overrightarrow{\texttt{mouse}}$

Polysemy

The distributional vector for a word (e.g. $\overrightarrow{\texttt{mouse}}$) encodes information on *all* occurrences of the word

If a word has several different meanings in a corpus (e.g. `mouse`), they will all be represented in the same distributional vector

Is this a problem?

Polysemy

The distributional vector for a word (e.g. $\overrightarrow{\texttt{mouse}}$) encodes information on *all* occurrences of the word

If a word has several different meanings in a corpus (e.g. `mouse`), they will all be represented in the same distributional vector

Is this a problem?

Polysemy

The distributional vector for a word (e.g. $\overrightarrow{\texttt{mouse}}$) encodes information on *all* occurrences of the word

If a word has several different meanings in a corpus (e.g. `mouse`), they will all be represented in the same distributional vector

Is this a problem?

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

Polysemy can be a problem when doing nearest neighbor analysis:

	General (BNC)
hot	boiling
	distilled
	brackish
	drinking
	cold
cold	hot
	franco-prussian
	boer
	iran-iraq
	napoleonic

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

Polysemy can be a problem when doing nearest neighbor analysis:

	General (BNC)
	boiling
	distilled
hot	brackish
	drinking
	cold
	hot
	franco-prussian
cold	boer
	iran-iraq
	napoleonic

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

Paradigmatic nearest neighbors share syntagmatic relations

Disambiguate the neighbors by sorting the syntagmatic relations

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

Paradigmatic nearest neighbors share syntagmatic relations

Disambiguate the neighbors by sorting the syntagmatic relations

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

1. Extract a word's k nearest paradigmatic neighbours
2. Extract the word's m nearest left and right syntagmatic neighbours ($m < k$)
3. For each of the k nearest paradigmatic neighbours, extract its m nearest left and right syntagmatic neighbours
   - If any of the m nearest left or right syntagmatic neighbours are equal to any of the target word's left or right syntagmatic neighbours, use that syntagmatic neighbour as a label for the paradigmatic relation
4. Nearest neighbours that have the same syntagmatic label belong to the same paradigm

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

1. Extract a word's k nearest paradigmatic neighbours
2. Extract the word's m nearest left and right syntagmatic neighbours ($m < k$)
3. For each of the k nearest paradigmatic neighbours, extract its m nearest left and right syntagmatic neighbours
   - If any of the m nearest left or right syntagmatic neighbours are equal to any of the target word's left or right syntagmatic neighbours, use that syntagmatic neighbour as a label for the paradigmatic relation
4. Nearest neighbours that have the same syntagmatic label belong to the same paradigm

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

1. Extract a word's k nearest paradigmatic neighbours
2. Extract the word's m nearest left and right syntagmatic neighbours ($m < k$)
3. For each of the k nearest paradigmatic neighbours, extract its m nearest left and right syntagmatic neighbours
   - If any of the m nearest left or right syntagmatic neighbours are equal to any of the target word's left or right syntagmatic neighbours, use that syntagmatic neighbour as a label for the paradigmatic relation
4. Nearest neighbours that have the same syntagmatic label belong to the same paradigm

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

1. Extract a word's k nearest paradigmatic neighbours
2. Extract the word's m nearest left and right syntagmatic neighbours ($m < k$)
3. For each of the k nearest paradigmatic neighbours, extract its m nearest left and right syntagmatic neighbours
   - If any of the m nearest left or right syntagmatic neighbours are equal to any of the target word's left or right syntagmatic neighbours, use that syntagmatic neighbour as a label for the paradigmatic relation
4. Nearest neighbours that have the same syntagmatic label belong to the same paradigm

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

1. Extract a word's k nearest paradigmatic neighbours
2. Extract the word's m nearest left and right syntagmatic neighbours ($m < k$)
3. For each of the k nearest paradigmatic neighbours, extract its m nearest left and right syntagmatic neighbours
   - If any of the m nearest left or right syntagmatic neighbours are equal to any of the target word's left or right syntagmatic neighbours, use that syntagmatic neighbour as a label for the paradigmatic relation
4. Nearest neighbours that have the same syntagmatic label belong to the same paradigm

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

cold

war 0.46	water 0.45	weather 0.20	air 0.18
boer napoleonic outbreak laws waging prisoners gulf biafran crimean phoney falklands punic peloponnesian undeclared vietnam tug world	hot soapy warm drinking boiling semer distilled icy cool bottled murky northumbrian shallow piped brackish polluted salty pail tap fresh tepid choppy muddy bucket treading unvented scummy drinkable salted	warm humid wet	hot warm cool polluted fresh chilly chill humid

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

hot

water 0.69	cold 0.26	air 0.24
semer drinking boiling distilled soapy bottled cold piped northumbrian brackish pail tepid polluted shallow warm choppy iced murky drinkable icy fresh turbid salted muddy filtered cool ionized lukewarm	icy	cold polluted warm fresh pollution cool clean pollute sunlit

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

warm

welcome 0.37	air 0.24	water 0.24	weather 0.18
outstay warmly well overstay	cool hot cold clean fresh chilly moist polluted balmy chill suck thin compressed wintry gulp	cool hot cold clean fresh polluted icy heat salty drinking boiling soapy shallow murky	cold wet dry wintry

Polysemy

Syntagmatically labelled partitioning (Koptjevskaja-Tamm & Sahlgren, 2012)

lukewarm

response 0.40	water 0.34	coffee 0.34	support 0.18
orienting elicit immune attentional enthusiastic eyeblink galvanic evoked chebyshev exaggerated	drinking tepid distilled boiling semer hot soapy bottled iced shallow brackish distil drinkable pour choppy piped northumbrian scummy murky pail cold unvented treading turbid bucket instantaneous foaming polluted	decaffeinated sip ersatz espresso instant sipping tea pour mug decaffeinate	enthusiastic wholehearted

Compositionality

The principle of compositionality

The meaning of a complex expression is a function of the meanings of its parts and of their syntactic mode of combination

The ingredients of compositionality (Partee 1984)

- *a theory of lexical meanings:* assigns meanings to the atomic parts (e.g. words)
- *a theory of syntactic structures:* determines the structure of complex expressions
- *a theory of semantic composition:* determines functions that compose meanings

Compositionality

The principle of compositionality

The meaning of a complex expression is a function of the meanings of its parts and of their syntactic mode of combination

The ingredients of compositionality (Partee 1984)

- *a theory of lexical meanings:* assigns meanings to the atomic parts (e.g. words)
- *a theory of syntactic structures:* determines the structure of complex expressions
- *a theory of semantic composition:* determines functions that compose meanings

Compositionality

The principle of compositionality

The meaning of a complex expression is a function of the meanings of its parts and of their syntactic mode of combination

The ingredients of compositionality (Partee 1984)

- *a theory of lexical meanings:* assigns meanings to the atomic parts (e.g. words)
- *a theory of syntactic structures:* determines the structure of complex expressions
- *a theory of semantic composition:* determines functions that compose meanings

Compositionality

The principle of compositionality

The meaning of a complex expression is a function of the meanings of its parts and of their syntactic mode of combination

The ingredients of compositionality (Partee 1984)

- *a theory of lexical meanings:* assigns meanings to the atomic parts (e.g. words)
- *a theory of syntactic structures:* determines the structure of complex expressions
- *a theory of semantic composition:* determines functions that compose meanings

Compositionality

1. What are the semantic operations that drive compositionality in word spaces?
2. What is the interpretation to be assigned to complex expressions (e.g. phrases, sentences, etc.) in word spaces?
3. How to represent the meaning of words in context?
4. How to account for the dependence of compositional meaning on syntactic structure?
5. How to test compositional representations in word spaces?
6. *Do we need do worry about compositionality?*

Compositionality

1. What are the semantic operations that drive compositionality in word spaces?
2. What is the interpretation to be assigned to complex expressions (e.g. phrases, sentences, etc.) in word spaces?
3. How to represent the meaning of words in context?
4. How to account for the dependence of compositional meaning on syntactic structure?
5. How to test compositional representations in word spaces?
6. *Do we need do worry about compositionality?*

Compositionality

1. What are the semantic operations that drive compositionality in word spaces?
2. What is the interpretation to be assigned to complex expressions (e.g. phrases, sentences, etc.) in word spaces?
3. How to represent the meaning of words in context?
4. How to account for the dependence of compositional meaning on syntactic structure?
5. How to test compositional representations in word spaces?
6. *Do we need do worry about compositionality?*

Compositionality

1. What are the semantic operations that drive compositionality in word spaces?
2. What is the interpretation to be assigned to complex expressions (e.g. phrases, sentences, etc.) in word spaces?
3. How to represent the meaning of words in context?
4. How to account for the dependence of compositional meaning on syntactic structure?
5. How to test compositional representations in word spaces?
6. *Do we need do worry about compositionality?*

Compositionality

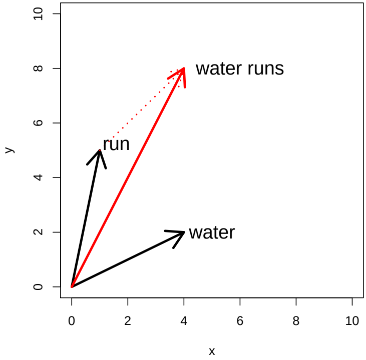
1. What are the semantic operations that drive compositionality in word spaces?
2. What is the interpretation to be assigned to complex expressions (e.g. phrases, sentences, etc.) in word spaces?
3. How to represent the meaning of words in context?
4. How to account for the dependence of compositional meaning on syntactic structure?
5. How to test compositional representations in word spaces?
6. *Do we need do worry about compositionality?*

Compositionality

1. What are the semantic operations that drive compositionality in word spaces?
2. What is the interpretation to be assigned to complex expressions (e.g. phrases, sentences, etc.) in word spaces?
3. How to represent the meaning of words in context?
4. How to account for the dependence of compositional meaning on syntactic structure?
5. How to test compositional representations in word spaces?
6. *Do we need do worry about compositionality?*

Compositionality

The distributional "meaning" of a phrase as the combined vector built with the vectors of the words in the phrase



Compositionality

Types of vector composition

Simple vector sum (Landauer & Dumais 1997)

- $\overrightarrow{p} = \overrightarrow{a} + \overrightarrow{b}$
- $\overrightarrow{chase\ cat} = \overrightarrow{chase} + \overrightarrow{cat}$

Compositionality

Types of vector composition

Simple vector sum (Landauer & Dumais 1997)

- $\overrightarrow{p} = \overrightarrow{a} + \overrightarrow{b}$
- $\overrightarrow{chase\ cat} = \overrightarrow{chase} + \overrightarrow{cat}$

Compositionality

Types of vector composition

Simple vector sum (Landauer & Dumais 1997)

- $\overrightarrow{p} = \overrightarrow{a} + \overrightarrow{b}$
- $\overrightarrow{chase\,cat} = \overrightarrow{chase} + \overrightarrow{cat}$

Compositionality

Types of vector composition

Context-sensitive vector sum (Kintsch 2001)

- $\overrightarrow{p} = \overrightarrow{a} + \overrightarrow{b} + \sum \overrightarrow{k}$
  where k are the k nearest neighbors of the predicate which are also neighbors of the argument
- $\overrightarrow{chase\,cat} = \overrightarrow{chase} + \overrightarrow{cat} + (\overrightarrow{hunt} + \overrightarrow{prey} + \ldots + \overrightarrow{capture})$

Compositionality

Types of vector composition

Context-sensitive vector sum (Kintsch 2001)

- $\overrightarrow{p} = \overrightarrow{a} + \overrightarrow{b} + \sum \overrightarrow{k}$
  where k are the k nearest neighbors of the predicate which are also neighbors of the argument
- $\overrightarrow{chase\,cat} = \overrightarrow{chase} + \overrightarrow{cat} + (\overrightarrow{hunt} + \overrightarrow{prey} + \ldots + \overrightarrow{capture})$

Compositionality

Types of vector composition

Context-sensitive vector sum (Kintsch 2001)

- $\overrightarrow{p} = \overrightarrow{a} + \overrightarrow{b} + \sum \overrightarrow{k}$
  where k are the k nearest neighbors of the predicate which are also neighbors of the argument
- $\overrightarrow{chase\ cat} = \overrightarrow{chase} + \overrightarrow{cat} + (\ \overrightarrow{hunt} + \overrightarrow{prey} + \ldots + \overrightarrow{capture})$

Compositionality

Types of vector composition

Context-sensitive vector sum (Kintsch 2001)

- $\overrightarrow{p} = \overrightarrow{a} + \overrightarrow{b} + \sum \overrightarrow{k}$
  where k are the k nearest neighbors of the predicate which are also neighbors of the argument
- $\overrightarrow{chase\ cat} = \overrightarrow{chase} + \overrightarrow{cat} + (\ \overrightarrow{hunt} + \overrightarrow{prey} + \ldots + \overrightarrow{capture})$

Compositionality

Types of vector composition

Pairwise multiplication (Mitchell & Lapata 2010)

- $\overrightarrow{p} = \overrightarrow{a} \cdot \overrightarrow{b}$
- $\overrightarrow{chase\ cat} = \overrightarrow{chase} \cdot \overrightarrow{cat}$

Compositionality

Types of vector composition

Pairwise multiplication (Mitchell & Lapata 2010)

- $\overrightarrow{p} = \overrightarrow{a} \cdot \overrightarrow{b}$
- $\overrightarrow{chase\ cat} = \overrightarrow{chase} \cdot \overrightarrow{cat}$

Compositionality

Types of vector composition

Pairwise multiplication (Mitchell & Lapata 2010)

- $\overrightarrow{p} = \overrightarrow{a} \cdot \overrightarrow{b}$
- $\overrightarrow{chase\ cat} = \overrightarrow{chase} \cdot \overrightarrow{cat}$

Compositionality

Additive composition preserves all the dimensions of the component vectors

	hacker	cheese	button
mouse	25	10	17
click	30	0	20
click mouse	55	10	37

Compositionality

Additive composition preserves all the dimensions of the component vectors

	hacker	cheese	button
mouse	25	10	17
click	30	0	20
click mouse	55	10	37

Compositionality

Multiplicative composition selects only the dimensions shared by the component vectors

	hacker	cheese	button
mouse	25	10	17
click	30	0	20
click mouse	1650	0	340

Compositionality

Multiplicative composition selects only the dimensions shared by the component vectors

	hacker	cheese	button
mouse	25	10	17
click	30	0	20
click mouse	1650	0	340

Compositionality

Other (more or less computationally demanding) approaches

- Tensor product
- Circular convolution

Compositionality

Other (more or less computationally demanding) approaches

- Tensor product
- Circular convolution

Compositionality

Other (more or less computationally demanding) approaches

- Tensor product
- Circular convolution

Compositionality

Applications

- Paraphrases
- Analogies
- IR

(Multiplicative models typically outperform additive models)

Compositionality

Applications

- Paraphrases
- Analogies
- IR

(Multiplicative models typically outperform additive models)

Compositionality

Applications

- Paraphrases
- Analogies
- IR

(Multiplicative models typically outperform additive models)

Compositionality

Applications

- Paraphrases
- Analogies
- IR

(Multiplicative models typically outperform additive models)

Compositionality

Applications

- Paraphrases
- Analogies
- IR

(Multiplicative models typically outperform additive models)

Semantic Relations

Research goal

To extend the ability of current word space models to discriminate among different types of semantic relations (e.g. coordinate, hypernyms, synonyms, antonyms, etc.)

Semantic Relations

Research goal

To extend the ability of current word space models to discriminate among different types of semantic relations (e.g. coordinate, hypernyms, synonyms, antonyms, etc.)

Semantic Relations

Research goal

To extend the ability of current word space models to discriminate among different types of semantic relations (e.g. coordinate, hypernyms, synonyms, antonyms, etc.)

Semantic Relations

Hierarchical structure has a central role in the organization of semantic memory (Murphy & Lassaline 1997)

A taxonomy is a sequence of progressively broader categories, related by the inclusion relation (ISA, *hypernymy*)

`golden retriever` $\subset$ `dog` $\subset$ `animal` $\subset$ `living thing` $\subset$ `physical entity` ...

Semantic Relations

Hierarchical structure has a central role in the organization of semantic memory (Murphy & Lassaline 1997)

A taxonomy is a sequence of progressively broader categories, related by the inclusion relation (ISA, *hypernymy*)

`golden retriever` $\subset$ `dog` $\subset$ `animal` $\subset$ `living thing` $\subset$ `physical entity` ...

Semantic Relations

Hierarchical structure has a central role in the organization of semantic memory (Murphy & Lassaline 1997)

A taxonomy is a sequence of progressively broader categories, related by the inclusion relation (ISA, *hypernymy*)

`golden retriever` $\subset$ `dog` $\subset$ `animal` $\subset$ `living thing` $\subset$ `physical entity` ...

Semantic Relations

Hierarchical structure has a central role in the organization of semantic memory (Murphy & Lassaline 1997)

A taxonomy is a sequence of progressively broader categories, related by the inclusion relation (ISA, *hypernymy*)

`golden retriever` $\subset$ `dog` $\subset$ `animal` $\subset$ `living thing` $\subset$ `physical entity` ...

Semantic Relations

The levels of a taxonomy represent different granularities on which an entity can be categorized *(they are paradigmatically related)*

A golden retriever, or a dog, or an animal,
or a living thing, or a physical entity...

Semantic networks (e.g. WordNet) are organized around taxonomical relations

Semantic Relations

The levels of a taxonomy represent different granularities on which an entity can be categorized *(they are paradigmatically related)*

A golden retriever, or a dog, or an animal, or a living thing, or a physical entity...

Semantic networks (e.g. WordNet) are organized around taxonomical relations

Semantic Relations

The levels of a taxonomy represent different granularities on which an entity can be categorized *(they are paradigmatically related)*

A golden retriever, or a dog, or an animal, or a living thing, or a physical entity...

Semantic networks (e.g. WordNet) are organized around taxonomical relations

Semantic Relations

Hypernymy is an *asymmetric* relation

- X is a dog $\Rightarrow$ X is an animal
- X is an animal $\not\Rightarrow$ X is a dog

Standard distributional similarity measures are symmetric

- $cosine(x, y) = cosine(y, x)$

Semantic Relations

Hypernymy is an *asymmetric* relation

- X is a dog $\Rightarrow$ X is an animal
- X is an animal $\not\Rightarrow$ X is a dog

Standard distributional similarity measures are symmetric

- $cosine(x, y) = cosine(y, x)$

Semantic Relations

Hypernymy is an *asymmetric* relation

- X is a dog $\Rightarrow X$ is an animal
- X is an animal $\nRightarrow X$ is a dog

Standard distributional similarity measures are symmetric

- $cosine(x, y) = cosine(y, x)$

Semantic Relations

Hypernymy is an *asymmetric* relation

- X is a dog $\Rightarrow$ X is an animal
- X is an animal $\not\Rightarrow$ X is a dog

Standard distributional similarity measures are symmetric

- $cosine(x, y) = cosine(y, x)$

Semantic Relations

Hypernymy is an *asymmetric* relation

- X is a dog $\Rightarrow$ X is an animal
- X is an animal $\nRightarrow$ X is a dog

Standard distributional similarity measures are symmetric

- $cosine(x, y) = cosine(y, x)$

Semantic Relations

Hypernyms are *semantically broader* than their hyponyms

Extensionally broader: `animal` refers to a broader set of entities than `dog`

Intensionally broader: `dog` has more specific properties (e.g. barking) than `animal`

Semantic Relations

Hypernyms are *semantically broader* than their hyponyms

Extensionally broader: `animal` refers to a broader set of entities than `dog`

Intensionally broader: `dog` has more specific properties (e.g. barking) than `animal`

Semantic Relations

Hypernyms are *semantically broader* than their hyponyms

Extensionally broader: `animal` refers to a broader set of entities than `dog`

Intensionally broader: `dog` has more specific properties (e.g. barking) than `animal`

Semantic Relations

Distributional inclusion hypothesis

if u is a semantically narrower term than v, then a significant number of salient distributional features of u is included in the feature vector of v as well, *but not the other way around*

Use *directional* similarity measures that are *asymmetric*

- $cosine(x, y) \neq cosine(y, x)$

Semantic Relations

Distributional inclusion hypothesis

if u is a semantically narrower term than v, then a significant number of salient distributional features of u is included in the feature vector of v as well, *but not the other way around*

Use *directional* similarity measures that are *asymmetric*

- $cosine(x, y) \neq cosine(y, x)$

Semantic Relations

Distributional inclusion hypothesis

if u is a semantically narrower term than v, then a significant number of salient distributional features of u is included in the feature vector of v as well, *but not the other way around*

Use *directional* similarity measures that are *asymmetric*

- $cosine(x, y) \neq cosine(y, x)$

Semantic Relations

Distributional inclusion hypothesis

if u is a semantically narrower term than v, then a significant number of salient distributional features of u is included in the feature vector of v as well, *but not the other way around*

Use *directional* similarity measures that are *asymmetric*

- $cosine(x, y) \neq cosine(y, x)$

Semantic Relations

- *WeedsPrec* (Weeds & Weir, 2003; Weeds et al., 2004)

$$WeedsPrec(u, v) = \frac{\sum_{f \in F_u \cap F_v} w_u(f)}{\sum_{f \in F_u} w_u(f)} \quad (1)$$

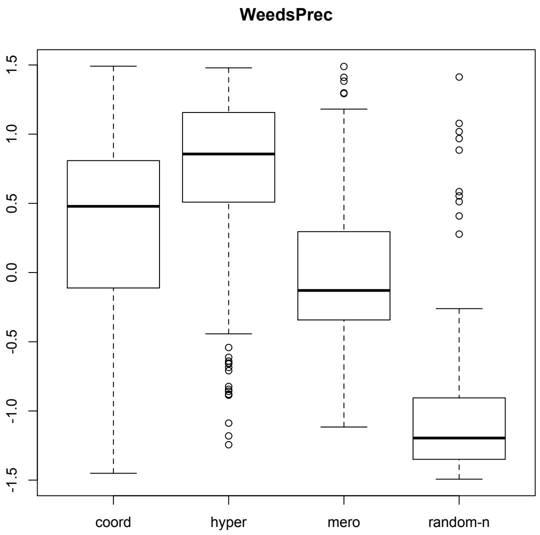
- *ClarkeDE* (Clarke 2009)

$$ClarkeDE(u, v) = \frac{\sum_{f \in F_u \cap F_v} min(w_u(f), w_v(f))}{\sum_{f \in F_u} w_u(f)} \quad (2)$$

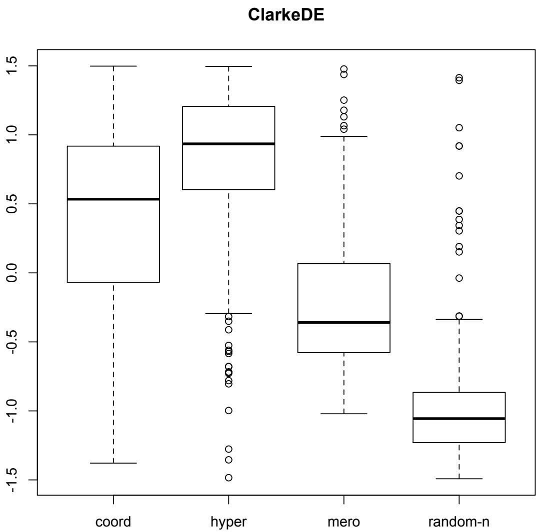
- *invCL* (Lenci & Benotto 2012)

$$invCL(u, v) = \sqrt{ClarkeDE(u, v) * (1 - ClarkeDE(v, u))} \quad (3)$$

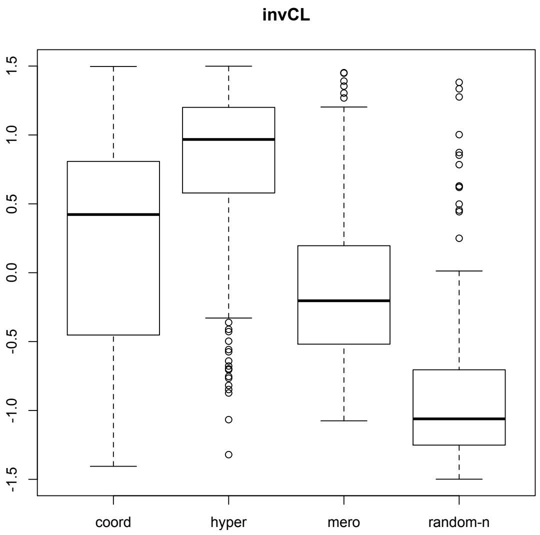
Directional Similarity Measures on BLESS



Directional Similarity Measures on BLESS



Directional Similarity Measures on BLESS



Visualization

How can we visualize the similarities?

Nearest neighbor lists

Visualization

How can we visualize the similarities?

Nearest neighbor lists

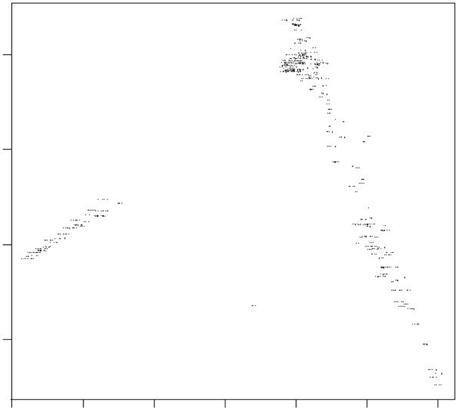
Visualization

How can we visualize the similarities?

Nearest neighbor lists

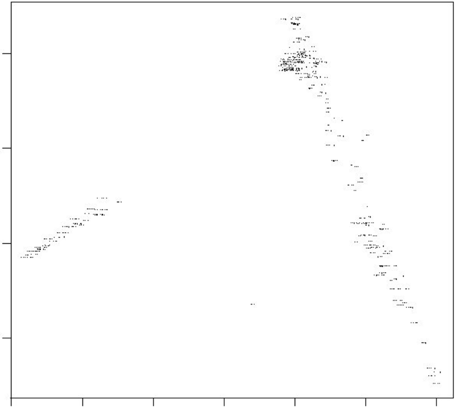
Visualization

Multidimensional scaling



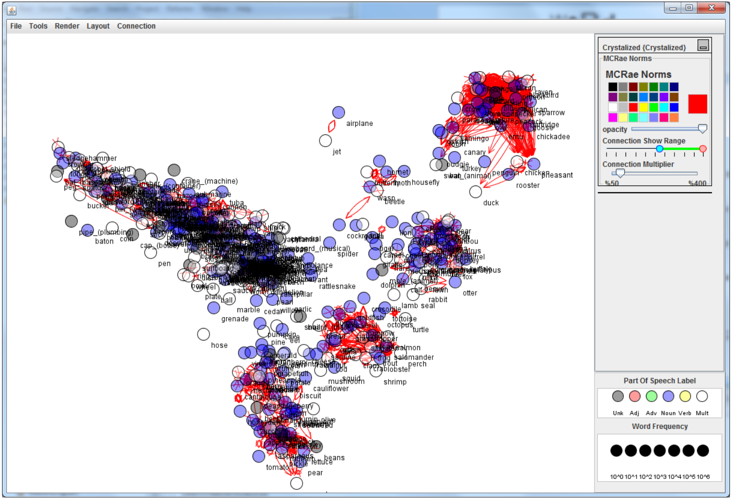
Visualization

Multidimensional scaling



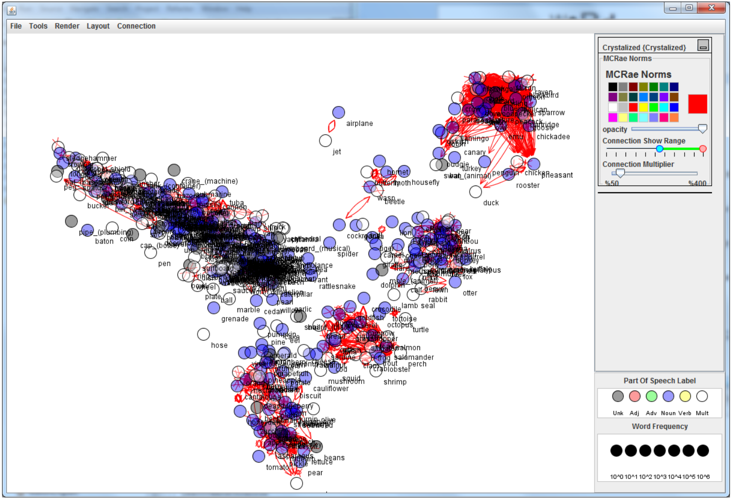
Visualization

Word 2 Word



Visualization

Word 2 Word



Lab

Use Word 2 Word to:

- Experiment with visualization of word spaces

Lab

Use Word 2 Word to:

- Experiment with visualization of word spaces

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda

Remember!

- The Distributional Hypothesis
- Syntagmatic, paradigmatic and topical similarities
- Co-occurrence matrix and distributional vectors
- Vector similarity and nearest neighbors
- Words-by-regions matrices and LSA
- Words-by-words-matrices and HAL
- Dependency-based models
- Random Indexing and Random Permutations
- (Problems of) evaluation
- (Examples of) tasks
- (The current) research agenda