utterance spaces — how to represent lexical items, constructions, and contextual data in a unified vector space

jussi karlgren

gavagai | kth

august 2018

Jussi Karlgren

- linguistics, stylistics, genre ↔ statistical models of language in use, distributional semantics
- mostly at Gavagai doing text analysis for commercial clients lexicon.gavagai.se
- adjoint professor of language technology at KTH Royal Institute of Technology, Stockholm and docent at Helsinki University
- (visitor at Stanford University Dept of Linguistics, 2017-18)

Jussi Karlgren

- linguistics, stylistics, genre ↔ statistical models of language in use, distributional semantics
- mostly at Gavagai doing text analysis for commercial clients lexicon.gavagai.se
- adjoint professor of language technology at KTH Royal Institute of Technology, Stockholm and docent at Helsinki University
- (visitor at Stanford University Dept of Linguistics, 2017-18)

Jussi Karlgren

- linguistics, stylistics, genre ↔ statistical models of language in use, distributional semantics
- mostly at Gavagai doing text analysis for commercial clients lexicon.gavagai.se
- adjoint professor of language technology at KTH Royal Institute of Technology, Stockholm and docent at Helsinki University
- (visitor at Stanford University Dept of Linguistics, 2017-18)

Jussi Karlgren

- linguistics, stylistics, genre ↔ statistical models of language in use, distributional semantics
- mostly at Gavagai doing text analysis for commercial clients lexicon.gavagai.se
- adjoint professor of language technology at KTH Royal Institute of Technology, Stockholm and docent at Helsinki University
- (visitor at Stanford University Dept of Linguistics, 2017-18)

how can we use our largely symbolic understanding of language structure in continuous processing models?

semantic spaces: a powerful metaphor and useful representation

- (relatively) low-dimensional projections of high-dimensional observational data
- very popular in recent approaches (aka embeddings)
- geometry is intuitively appealing and plausible
- manageable implementational framework

- Gerard Salton et al. 1975. A vector space model for automatic indexing. CACM
- Hinrich Schütze. 1993. Word Space. NIPS.
- David Dubin. 2004. „The most influential paper Gerard Salton never wrote". Library Trends.
- Magnus Sahlgren. 2006. The Word-Space Model. Stockholm University.

semantic spaces: a powerful metaphor and useful representation

- (relatively) low-dimensional projections of high-dimensional observational data
- very popular in recent approaches (aka embeddings)
- geometry is intuitively appealing and plausible
- manageable implementational framework

- Gerard Salton et al. 1975. A vector space model for automatic indexing. CACM
- Hinrich Schütze. 1993. Word Space. NIPS.
- David Dubin. 2004. „The most influential paper Gerard Salton never wrote". Library Trends.
- Magnus Sahlgren. 2006. The Word-Space Model. Stockholm University.

semantic spaces: a powerful metaphor and useful representation

- (relatively) low-dimensional projections of high-dimensional observational data
- very popular in recent approaches (aka embeddings)
- geometry is intuitively appealing and plausible
- manageable implementational framework

- Gerard Salton et al. 1975. A vector space model for automatic indexing. CACM
- Hinrich Schütze. 1993. Word Space. NIPS.
- David Dubin. 2004. „The most influential paper Gerard Salton never wrote". Library Trends.
- Magnus Sahlgren. 2006. The Word-Space Model. Stockholm University.

semantic spaces: a powerful metaphor and useful representation

- (relatively) low-dimensional projections of high-dimensional observational data
- very popular in recent approaches (aka embeddings)
- geometry is intuitively appealing and plausible
- manageable implementational framework

$$d_{cos}(x, y) = \frac{\vec{x} \cdot \vec{y}}{\mid \vec{x} \mid\mid \vec{y} \mid} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Gerard Salton et al. 1975. A vector space model for automatic indexing. CACM
- Hinrich Schütze. 1993. Word Space. NIPS.
- David Dubin. 2004. „The most influential paper Gerard Salton never wrote". Library Trends.
- Magnus Sahlgren. 2006. The Word-Space Model. Stockholm University.

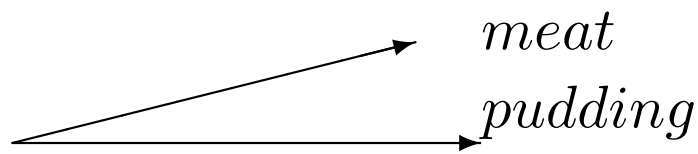
we have spatial intuitions which guide our understanding of semantic spaces

we are prepared to jettison much of symbolic processing to achieve this but still look for symbolic meaning in the space

woman vs man ≈ queen vs king

But

hallucinated precision beyond the semantic horizon



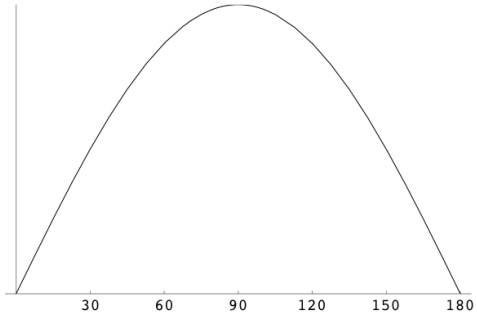
$d_{cos}(\text{"}pudding\text{"}, \text{"}meat\text{"}) \approx cos(\pi/5)$
$d_{cos}(\text{"}cardamom\text{"}, \text{"}tensor\ \ algebra\text{"}) =?$

- Jussi Karlgren. 2005. "Meaningful models for information access systems." CSLI

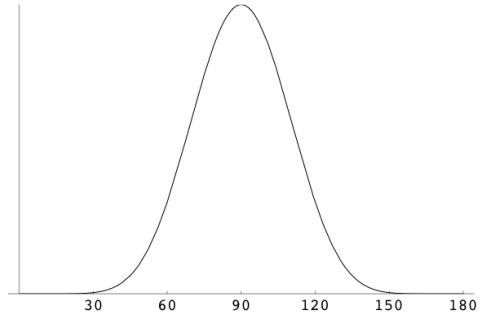
expected angle distribution

dimensionality and distance metrics render centroids less reliable than one believes probability

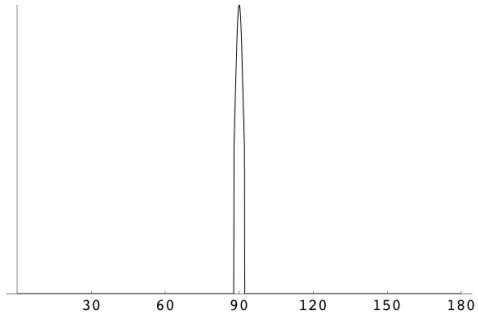
distribution for angles between directions to random points in many-dimensional spaces.



3 dimensions



10 dimensions



1000 dimensions

- Jussi Karlgren et al. 2008. Filaments of Meaning in Word Space. ECIR

still, a good thing

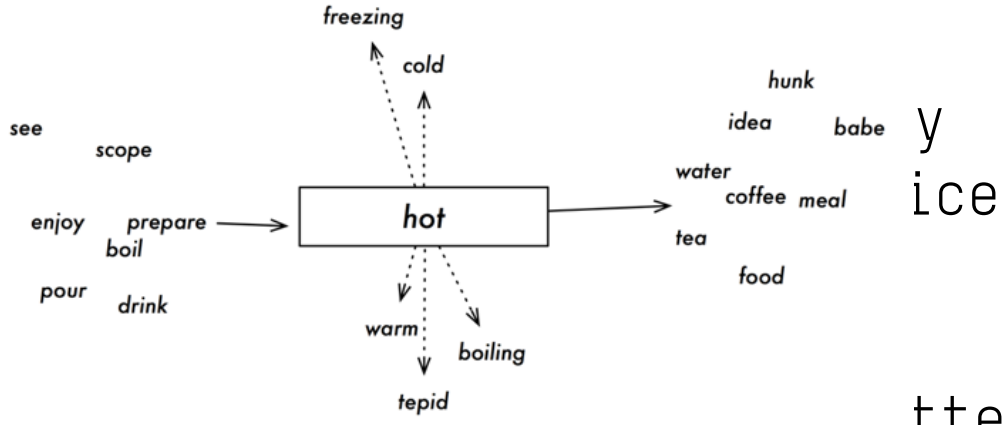
distributional semantics: populating semantic spaces

the weather is great in barcelona
the weather is gray in stockholm
the weather is hot in évora
the climate is passable in nice
the weather is chilly in helsinki
the weather is nippy in moscow
the weather is nice in hong kong
the weather in syktyvkar is balmy
the climate is chilly at the office
the tea is hot
i drink tea
a hot meal will make you feel better
enjoy your hot beverages

- Zellig Harris. 1968. Mathematical structures of language.

distributional semantics: populating semantic spaces

the weather is great in barcelona
the weather is gray in stockholm
the weather is hot in évora
the climate is passable in nice
the weather is chilly in helsinki
the weather is nippy in moscow
the
the
the
the
i dr
a hc
enjoy your hot beverages

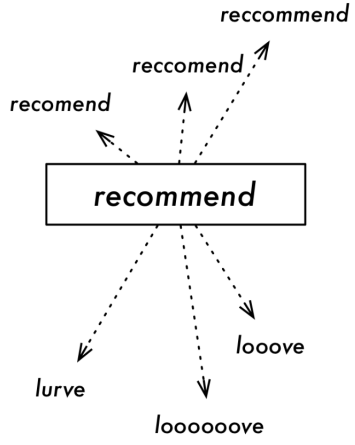


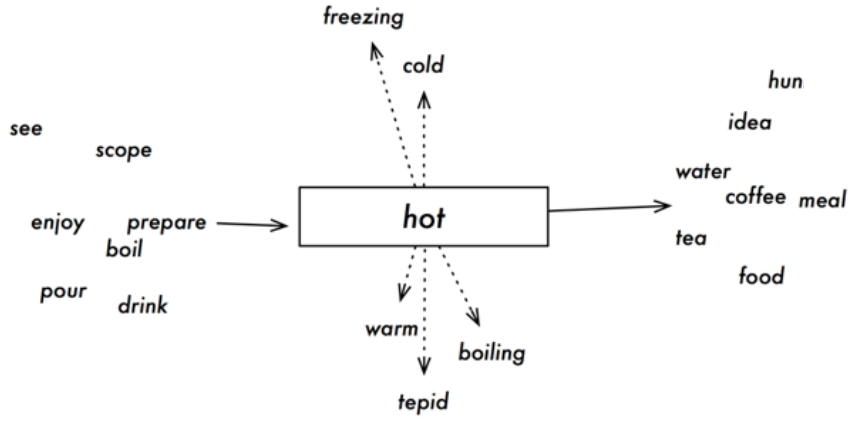
ice
tter

- Zellig Harris. 1968. Mathematical structures of language.

distributional semantics: populating semantic spaces

the weather is great in barcelona
the weather is gray in stockholm
the weather is hot in évora
the climate is passable in ni
the weather is chilly in hels
the weather is nippy in moscc
the
the
the
the
i dr
a hc
enjoy your hot beverages





ice

tter

- Zellig Harris. 1968. Mathematical structures of language.

constructional items in the linguistic signal

fairly strong claim:

the pattern of an utterance is a feature with the same ontological status as the terms that occur in the utterance

constructions and lexemes both have conceptual meaning

there is no solid argument to draw a line between them

- Naomi Sager et al. 1965-1998. Linguistic String Project.
- William Croft. 2005. Radical and typological arguments for radical construction grammar. In Construction Grammars John Benjamins.

constructional items in the linguistic signal

fairly strong claim:

the pattern of an utterance is a feature with the same ontological status as the terms that occur in the utterance

constructions and lexemes both have conceptual meaning

there is no solid argument to draw a line between them

- Naomi Sager et al. 1965-1998. Linguistic String Project.
- William Croft. 2005. Radical and typological arguments for radical construction grammar. In Construction Grammars John Benjamins.

constructional items in the linguistic signal

fairly strong claim:

the pattern of an utterance is a feature with the same ontological status as the terms that occur in the utterance

constructions and lexemes both have conceptual meaning

there is no solid argument to draw a line between them

- Naomi Sager et al. 1965-1998. Linguistic String Project.
- William Croft. 2005. Radical and typological arguments for radical construction grammar. In Construction Grammars John Benjamins.

we will use this approach

hyperdimensional computing: a powerful and principled processing model

- fixed dimensional index vectors or labels for basic features
- aggregated context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- allows explicit feature engineering if desirable
- obviates need for dimensionality reduction

- Tony Plate. 1995. "Holographic reduced representations" IEEE Transactions on Neural Networks
- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

hyperdimensional computing: a powerful and principled processing model

- fixed dimensional index vectors or labels for basic features
- aggregated context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- allows explicit feature engineering if desirable
- obviates need for dimensionality reduction

- Tony Plate. 1995. "Holographic reduced representations" IEEE Transactions on Neural Networks
- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

hyperdimensional computing: a powerful and principled processing model

- fixed dimensional index vectors or labels for basic features
- aggregated context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- allows explicit feature engineering if desirable
- obviates need for dimensionality reduction

- Tony Plate. 1995. "Holographic reduced representations" IEEE Transactions on Neural Networks
- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

hyperdimensional computing: a powerful and principled processing model

- fixed dimensional index vectors or labels for basic features
- aggregated context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- allows explicit feature engineering if desirable
- obviates need for dimensionality reduction

- Tony Plate. 1995. "Holographic reduced representations" IEEE Transactions on Neural Networks
- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

hyperdimensional computing: a powerful and principled processing model

- fixed dimensional index vectors or labels for basic features
- aggregated context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- allows explicit feature engineering if desirable
- obviates need for dimensionality reduction

- Tony Plate. 1995. "Holographic reduced representations" IEEE Transactions on Neural Networks
- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

hyperdimensional operation: addition

- vector addition yields a vector similar to its inputs

- $\bar{v}_{dog} + \bar{v}_{chew} + \bar{v}_{bone} \sim \bar{v}_{bone}$

hyperdimensional operation: multiplication

- vector multiplication yields a vector dissimilar to its inputs but preserves similarity

- $A * B \not\sim A$
- $d_{cos}(Q * A, Q * B) = d_{cos}(A, B)$
- $A * A = 1$

useful for e.g. variable binding

hyperdimensional operation: multiplication

- vector multiplication yields a vector dissimilar to its inputs but preserves similarity

- $A * B \not\sim A$
- $d_{cos}(Q * A, Q * B) = d_{cos}(A, B)$
- $A * A = 1$

useful for e.g. variable binding

hyperdimensional operation: multiplication

- vector multiplication yields a vector dissimilar to its inputs but preserves similarity

- $A * B \not\sim A$
- $d_{cos}(Q * A, Q * B) = d_{cos}(A, B)$
- $A * A = 1$

useful for e.g. variable binding

hyperdimensional operation: permutation

- vector permutation allows multiple space relationships to be represented in the same spaces, allowing for e.g. sequences or even tensors

„Dogs chew bones."

- $\bar{v}_{dcb} = \bar{v}_{dog} + \bar{v}_{chew} + \bar{v}_{bone} + \Pi_{subject}(\bar{v}_{dog}) + \bar{v}_{tense} * \bar{v}_{present} + \ldots$

hyperdimensional operation: combing

- vector combing removes signal within a noise corridor to yield a sparse and more compact vector

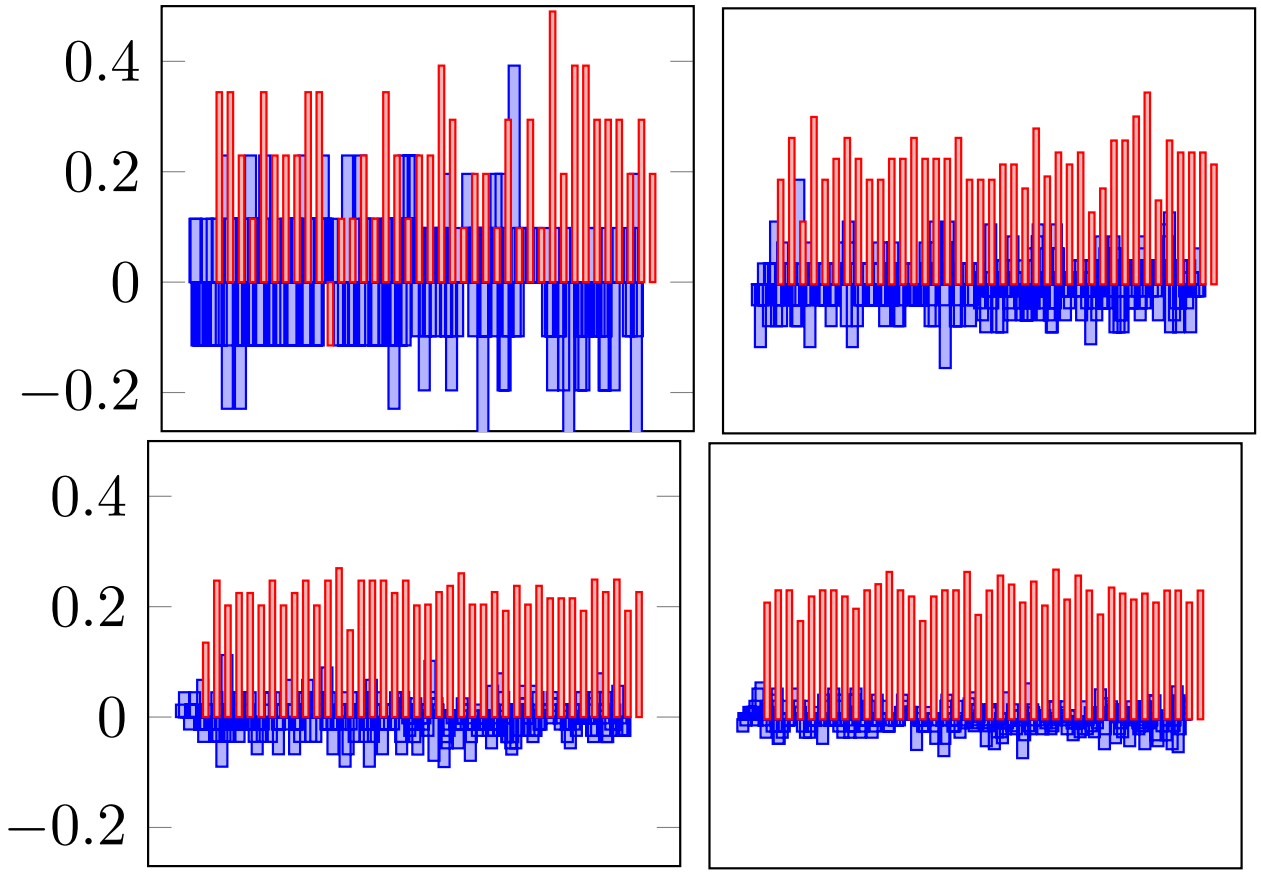


Figure 1: Cosine of feature vectors to state vector compared to random unrelated vectors in 100, 500, 1,000, and 2,000 dimensions

how then to populate a distributional vector space?

- „randomness is the path of least assumption"
- random indexing: how to bridge symbols and continuous models
- each item of interest is given a d-dimensional label with k non-zero cells randomly distributed over it
- each item of interest is also given an empty d-dimensional context vector in which distributional information is accrued

how then to populate a distributional vector space?

- „randomness is the path of least assumption”
- random indexing: how to bridge symbols and continuous models
- each item of interest is given a d-dimensional label with k non-zero cells randomly distributed over it
- each item of interest is also given an empty d-dimensional context vector in which distributional information is accrued

how then to populate a distributional vector space?

- „randomness is the path of least assumption”
- random indexing: how to bridge symbols and continuous models
- each item of interest is given a d-dimensional label with k non-zero cells randomly distributed over it
- each item of interest is also given an empty d-dimensional context vector in which distributional information is accrued

how then to populate a distributional vector space?

- „randomness is the path of least assumption”
- random indexing: how to bridge symbols and continuous models
- each item of interest is given a d-dimensional label with k non-zero cells randomly distributed over it
- each item of interest is also given an empty d-dimensional context vector in which distributional information is accrued

how then to populate a distributional vector space?

- „randomness is the path of least assumption”
- random indexing: how to bridge symbols and continuous models
- each item of interest is given a d-dimensional label with k non-zero cells randomly distributed over it
- each item of interest is also given an empty d-dimensional context vector in which distributional information is accrued

$d = 6; k = 2$		$d_{cos}(hot, chilly)$ 1.0
item	label vector	context vector
hot	...	[0, 0, 0, 0, 0, 0]
chilly	...	[0, 0, 0, 0, 0, 0]

the weather is hot ... the weather is chilly ...

$d = 6; k = 2$		$d_{cos}(hot, chilly)$ 1.0
item	label vector	context vector
hot	...	[0, -1, 0, 1, 0, 0]
chilly	...	[0, -1, 0, 1, 0, 0]
weather	[0, -1, 0, 1, 0, 0]	

the weather is hot ... the weather is chilly ...
the climate is hot ... the climate is chilly ...

$d = 6; k = 2$		$d_{cos}(hot, chilly)$ 1.0
item	label vector	context vector
hot	...	[1, -1, -1, 1, 0, 0]
chilly	...	[1, -1, -1, 1, 0, 0]
weather	[0, -1, 0, 1, 0, 0]	
climate	[1, 0, -1, 0, 0, 0]	

the weather is hot ... the weather is chilly ...
the climate is hot ... the climate is chilly ...
the weather pattern is hot ... the day turned chilly ...

$d = 6; k = 2$		$d_{cos}(hot, chilly)$ 0.47
item	label vector	context vector
hot	...	[1, -2, -1, 2, 0, 0]
chilly	...	[0, -1, -1, 0, 1, 1]
weather	[0, -1, 0, 1, 0, 0]	
climate	[1, 0, -1, 0, 0, 0]	
day	[0, 0, 0, -1, 1, 0]	
pattern	[0, 1, 0, 0, 0, -1]	
turned	[-1, 0, 0, 0, 0, 1]	

the weather is hot ... the weather is chilly ...
the climate is hot ... the climate is chilly ...
the weather pattern is hot ... the day turned chilly ...
the weather turned hot ... the climate is chilly ...

$d = 6; k = 2$	$d_{cos}(hot, chilly)$ 0.61	
item	label vector	context vector
hot	...	[1, -1, 0, 0, 0, -1]
chilly	...	[1, -1, -2, 0, 1, 1]
weather	[0, -1, 0, 1, 0, 0]	
climate	[1, 0, -1, 0, 0, 0]	
day	[0, 0, 0, -1, 1, 0]	
pattern	[0, 1, 0, 0, 0, -1]	
turned	[-1, 0, 0, 0, 0, 1]	

now do the same thing, but add in constructional elements!

now do the same thing, but add in constructional elements!

(a) semantic roles:
$\Pi_{agent}(\bar{v}_{dog}) + \Pi_{tense}(\bar{v}_{present}) + ...$

now do the same thing, but add in constructional elements!

(a) semantic roles:

$\Pi_{agent}(\bar{v}_{dog}) + \Pi_{tense}(\bar{v}_{present}) + ...$

(b) constructional elements:

$\bar{v}_{uncertain} + \bar{v}_{profanity} + \bar{v}_{utteranceverb} \cdots$

loads of sentences, represented as sums of features ($d = 2000; k = 10$):

loads of sentences, represented as sums of features ($d = 2000; k = 10$):

(a) words

loads of sentences, represented as sums of features ($d = 2000; k = 10$):

(a) words

(b) constructional elements such as above

loads of sentences, represented as sums of features ($d = 2000; k = 10$):

(a) words

(b) constructional elements such as above

find neighbours to:

„I really did not like the clarinet, I am afraid: it sounded weak!"

loads of sentences, represented as sums of features ($d = 2000; k = 10$):

(a) words

(b) constructional elements such as above

find neighbours to:

„I really did not like the clarinet, I am afraid: it sounded weak!"

(a) words: My sister plays the clarinet.

loads of sentences, represented as sums of features ($d = 2000; k = 10$):

(a) words

(b) constructional elements such as above

find neighbours to:

„I really did not like the clarinet, I am afraid: it sounded weak!"

(a) words: My sister plays the clarinet.

(b) constructions: I'm surrounded by really soft decadent pillows which do not work for me at all.

loads of sentences, represented as sums of features ($d = 2000; k = 10$):

(a) words

(b) constructional elements such as above

find neighbours to:

„I really did not like the clarinet, I am afraid: it sounded weak!"

(a) words: My sister plays the clarinet.

(b) constructions: I'm surrounded by really soft decadent pillows which do not work for me at all.

(„i quit drinking coffee")

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

steps to be tested and evaluated by

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

steps to be tested and evaluated by theoretical soundness

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

steps to be tested and evaluated by theoretical soundness and benchmarking metrics,

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

steps to be tested and evaluated by theoretical soundness and benchmarking metrics, both on own merits and on utility for downstream task

work flow to develop new methods for understanding human language

1. identify a need for doing something which may be expressed in human language
2. think about how human language expresses such things
3. develop a method for extraction expressions of relevance from stream of human language
4. develop a method for computing the prevalence and strength of signal
5. design a way to display and show signal to stakeholders

steps to be tested and evaluated by theoretical soundness and benchmarking metrics, both on own merits and on utility for downstream task; entire approach to be validated separately from benchmarking

recap

- semantic spaces are convenient but overintuitive
- constructions are linguistic items
- all linguistic items can be represented in a common distributional representation
- hyperdimensional representations do not need dimensionality reduction
- random patterns are more convenient than localist representations
- this provides a bridge from symbolic to continous models

recap

- semantic spaces are convenient but overintuitive
- constructions are linguistic items
- all linguistic items can be represented in a common distributional representation
- hyperdimensional representations do not need dimensionality reduction
- random patterns are more convenient than localist representations
- this provides a bridge from symbolic to continous models

recap

- semantic spaces are convenient but overintuitive
- constructions are linguistic items
- all linguistic items can be represented in a common distributional representation
- hyperdimensional representations do not need dimensionality reduction
- random patterns are more convenient than localist representations
- this provides a bridge from symbolic to continous models

recap

- semantic spaces are convenient but overintuitive
- constructions are linguistic items
- all linguistic items can be represented in a common distributional representation
- hyperdimensional representations do not need dimensionality reduction
- random patterns are more convenient than localist representations
- this provides a bridge from symbolic to continous models

recap

- semantic spaces are convenient but overintuitive
- constructions are linguistic items
- all linguistic items can be represented in a common distributional representation
- hyperdimensional representations do not need dimensionality reduction
- random patterns are more convenient than localist representations
- this provides a bridge from symbolic to continous models

recap

- semantic spaces are convenient but overintuitive
- constructions are linguistic items
- all linguistic items can be represented in a common distributional representation
- hyperdimensional representations do not need dimensionality reduction
- random patterns are more convenient than localist representations
- this provides a bridge from symbolic to continous models

recap

- semantic spaces are convenient but overintuitive
- constructions are linguistic items
- all linguistic items can be represented in a common distributional representation
- hyperdimensional representations do not need dimensionality reduction
- random patterns are more convenient than localist representations
- this provides a bridge from symbolic to continous models