# how to build a learning lexicon from streaming text and how it might be useful

jussi karlgren

gavagai

june 2018

This work has been supported by Vinnova, the Swedish Governmental Agency for Innovation Systems under a Vinnmer Marie Curie Grant.

# Jussi Karlgren

- stylistics, genre, statistical models of language in use, distributional semantics
- founder and researcher at the text analysis company Gavagai
- docent of language technology at Helsinki University and adjoint professor of language technology at KTH Royal Institute of Technology, Stockholm
- (visitor at Stanford University Dept of Linguistics, 2017-18)

# Jussi Karlgren

- stylistics, genre, statistical models of language in use, distributional semantics
- founder and researcher at the text analysis company Gavagai
- docent of language technology at Helsinki University and adjoint professor of language technology at KTH Royal Institute of Technology, Stockholm
- (visitor at Stanford University Dept of Linguistics, 2017-18)

# Jussi Karlgren

- stylistics, genre, statistical models of language in use, distributional semantics
- founder and researcher at the text analysis company Gavagai
- docent of language technology at Helsinki University and adjoint professor of language technology at KTH Royal Institute of Technology, Stockholm
- (visitor at Stanford University Dept of Linguistics, 2017-18)

# Jussi Karlgren

- stylistics, genre, statistical models of language in use, distributional semantics
- founder and researcher at the text analysis company Gavagai
- docent of language technology at Helsinki University and adjoint professor of language technology at KTH Royal Institute of Technology, Stockholm
- (visitor at Stanford University Dept of Linguistics, 2017-18)

# Gavagai

- Based in Stockholm (with roots at SICS, Swedish Institute for Computer Science)
- Research in processing human language since late 1990's
- Now a text analysis company with services in 45 languages, processing millions of documents daily
  - customer-oriented text clustering by theme and sentiment explorer.gavagai.se
  - media monitoring monitor.gavagai.se
  - Gavagai living lexicon lexicon.gavagai.se

---


- Jussi Karlgren and Pentti Kanerva. 2018. "Hyperdimensional Utterance Spaces". DESIRE
- Fredrik Espinoza et al. 2018. "Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction." ACM CHIIR
- Magnus Sahlgren et al. 2017. "Gavagai Living Lexicon". LREC.
- Jussi Karlgren et al. 2012. "Usefulness of sentiment analysis." ECIR
- Magnus Sahlgren and Jussi Karlgren. 2009. "Terminology mining in social media." CIKM
- Pentti Kanerva et al. 2001. "Computing with large random patterns." CSLI

# Gavagai

- Based in Stockholm (with roots at SICS, Swedish Institute for Computer Science)
- Research in processing human language since late 1990's
- Now a text analysis company with services in 45 languages, processing millions of documents daily
  - customer-oriented text clustering by theme and sentiment explorer.gavagai.se
  - media monitoring monitor.gavagai.se
  - Gavagai living lexicon lexicon.gavagai.se

---


- Jussi Karlgren and Pentti Kanerva. 2018. "Hyperdimensional Utterance Spaces". DESIRE
- Fredrik Espinoza et al. 2018. "Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction." ACM CHIIR
- Magnus Sahlgren et al. 2017. "Gavagai Living Lexicon". LREC.
- Jussi Karlgren et al. 2012. "Usefulness of sentiment analysis." ECIR
- Magnus Sahlgren and Jussi Karlgren. 2009. "Terminology mining in social media." CIKM
- Pentti Kanerva et al. 2001. "Computing with large random patterns." CSLI

# Gavagai

- Based in Stockholm (with roots at SICS, Swedish Institute for Computer Science)
- Research in processing human language since late 1990's
- Now a text analysis company with services in 45 languages, processing millions of documents daily
  - customer-oriented text clustering by theme and sentiment explorer.gavagai.se
  - media monitoring monitor.gavagai.se
  - Gavagai living lexicon lexicon.gavagai.se

---


- Jussi Karlgren and Pentti Kanerva. 2018. "Hyperdimensional Utterance Spaces". DESIRE
- Fredrik Espinoza et al. 2018. "Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction." ACM CHIIR
- Magnus Sahlgren et al. 2017. "Gavagai Living Lexicon". LREC.
- Jussi Karlgren et al. 2012. "Usefulness of sentiment analysis." ECIR
- Magnus Sahlgren and Jussi Karlgren. 2009. "Terminology mining in social media." CIKM
- Pentti Kanerva et al. 2001. "Computing with large random patterns." CSLI

# Gavagai

- Based in Stockholm (with roots at SICS, Swedish Institute for Computer Science)
- Research in processing human language since late 1990's
- Now a text analysis company with services in 45 languages, processing millions of documents daily
  - customer-oriented text clustering by theme and sentiment explorer.gavagai.se
  - media monitoring monitor.gavagai.se
  - Gavagai living lexicon lexicon.gavagai.se

---


- Jussi Karlgren and Pentti Kanerva. 2018. "Hyperdimensional Utterance Spaces". DESIRE
- Fredrik Espinoza et al. 2018. "Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction." ACM CHIIR
- Magnus Sahlgren et al. 2017. "Gavagai Living Lexicon". LREC.
- Jussi Karlgren et al. 2012. "Usefulness of sentiment analysis." ECIR
- Magnus Sahlgren and Jussi Karlgren. 2009. "Terminology mining in social media." CIKM
- Pentti Kanerva et al. 2001. "Computing with large random patterns." CSLI

# Gavagai

- Based in Stockholm (with roots at SICS, Swedish Institute for Computer Science)
- Research in processing human language since late 1990's
- Now a text analysis company with services in 45 languages, processing millions of documents daily
  - customer-oriented text clustering by theme and sentiment explorer.gavagai.se
  - media monitoring monitor.gavagai.se
  - Gavagai living lexicon lexicon.gavagai.se

---


- Jussi Karlgren and Pentti Kanerva. 2018. "Hyperdimensional Utterance Spaces". DESIRE
- Fredrik Espinoza et al. 2018. "Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction." ACM CHIIR
- Magnus Sahlgren et al. 2017. "Gavagai Living Lexicon". LREC.
- Jussi Karlgren et al. 2012. "Usefulness of sentiment analysis." ECIR
- Magnus Sahlgren and Jussi Karlgren. 2009. "Terminology mining in social media." CIKM
- Pentti Kanerva et al. 2001. "Computing with large random patterns." CSLI

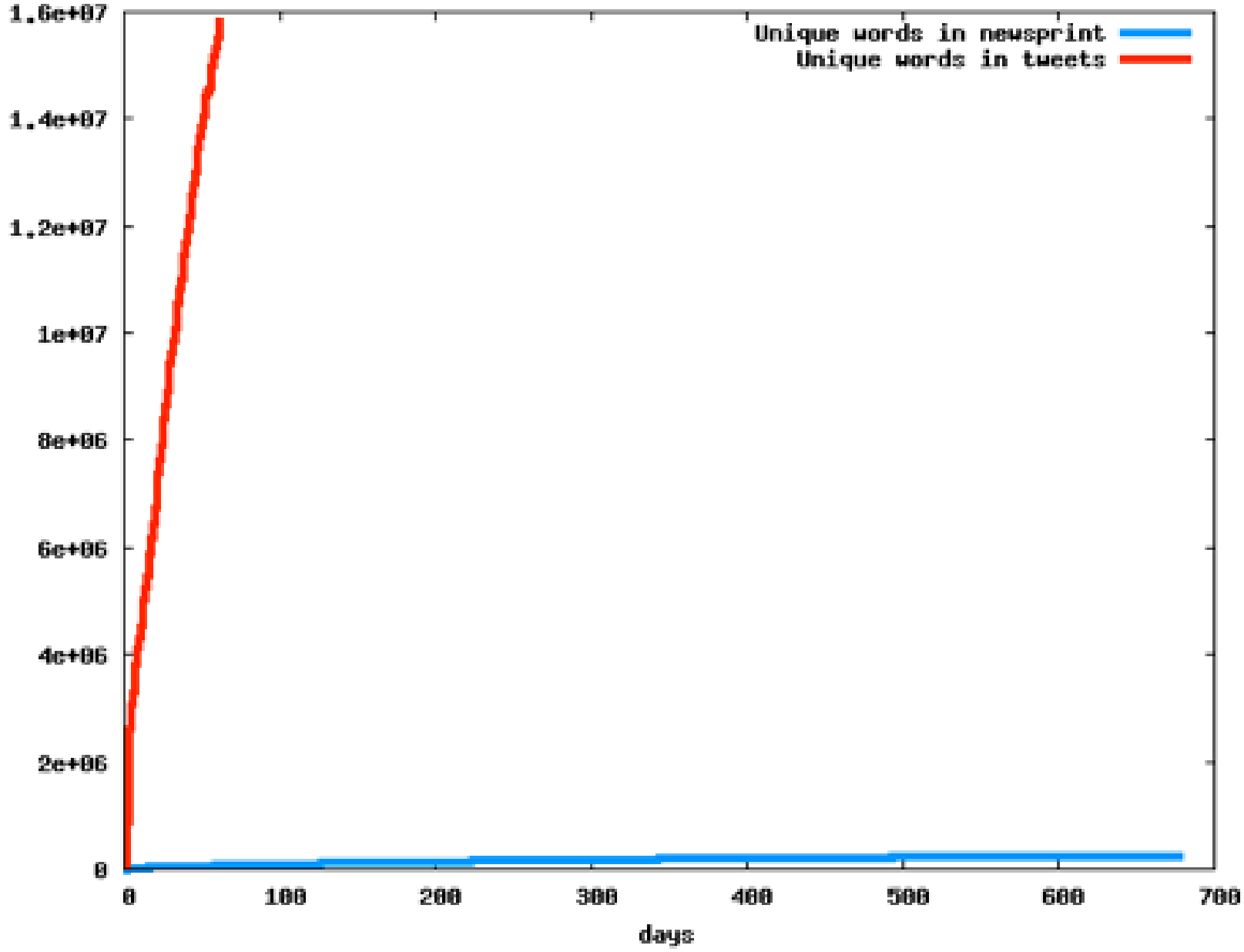
# Gavagai

- Based in Stockholm (with roots at SICS, Swedish Institute for Computer Science)
- Research in processing human language since late 1990's
- Now a text analysis company with services in 45 languages, processing millions of documents daily
  - customer-oriented text clustering by theme and sentiment explorer.gavagai.se
  - media monitoring monitor.gavagai.se
  - Gavagai living lexicon lexicon.gavagai.se

---


- Jussi Karlgren and Pentti Kanerva. 2018. "Hyperdimensional Utterance Spaces". DESIRE
- Fredrik Espinoza et al. 2018. "Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction." ACM CHIIR
- Magnus Sahlgren et al. 2017. "Gavagai Living Lexicon". LREC.
- Jussi Karlgren et al. 2012. "Usefulness of sentiment analysis." ECIR
- Magnus Sahlgren and Jussi Karlgren. 2009. "Terminology mining in social media." CIKM
- Pentti Kanerva et al. 2001. "Computing with large random patterns." CSLI

# language is a pilot case for high volume and high variety data



# human information processing

- human information processing is effective for streaming data
- handles analogy & saliency
- observes patterns and change rather than the literal
- operates with self-learning rather than instruction

# human information processing

- human information processing is effective for streaming data
- handles analogy & saliency
- observes patterns and change rather than the literal
- operates with self-learning rather than instruction

# human information processing

- human information processing is effective for streaming data
- handles analogy & saliency
- observes patterns and change rather than the literal
- operates with self-learning rather than instruction

# human information processing

- human information processing is effective for streaming data
- handles analogy & saliency
- observes patterns and change rather than the literal
- operates with self-learning rather than instruction

# human information processing

- human information processing is effective for streaming data
- handles analogy & saliency
- observes patterns and change rather than the literal
- operates with self-learning rather than instruction

worth keeping in mind as a model

# requirements for a knowledge representation

a representation should:

- have descriptive and explanatory power

  (allow backtracking into observations)

- be practical and convenient for further application

  (retain feature structure)

- be reasonably true to human performance

  (handle streaming and analogy!)

- handle patchy data

  (provide support for generalisation, defaults and constraints)

- be computationally habitable

  (not grow superlinearly with data input)

- be general

  (not tightly bound to some task)

# requirements for a knowledge representation

a representation should:

- have descriptive and explanatory power

  (allow backtracking into observations)

- be practical and convenient for further application

  (retain feature structure)

- be reasonably true to human performance

  (handle streaming and analogy!)

- handle patchy data

  (provide support for generalisation, defaults and constraints)

- be computationally habitable

  (not grow superlinearly with data input)

- be general

  (not tightly bound to some task)

# requirements for a knowledge representation

a representation should:

- have descriptive and explanatory power

  (allow backtracking into observations)

- be practical and convenient for further application

  (retain feature structure)

- be reasonably true to human performance

  (handle streaming and analogy!)

- handle patchy data

  (provide support for generalisation, defaults and constraints)

- be computationally habitable

  (not grow superlinearly with data input)

- be general

  (not tightly bound to some task)

# requirements for a knowledge representation

a representation should:

- have descriptive and explanatory power

  (allow backtracking into observations)

- be practical and convenient for further application

  (retain feature structure)

- be reasonably true to human performance

  (handle streaming and analogy!)

- handle patchy data

  (provide support for generalisation, defaults and constraints)

- be computationally habitable

  (not grow superlinearly with data input)

- be general

  (not tightly bound to some task)

# requirements for a knowledge representation

a representation should:

- have descriptive and explanatory power

  (allow backtracking into observations)

- be practical and convenient for further application

  (retain feature structure)

- be reasonably true to human performance

  (handle streaming and analogy!)

- handle patchy data

  (provide support for generalisation, defaults and constraints)

- be computationally habitable

  (not grow superlinearly with data input)

- be general

  (not tightly bound to some task)

# requirements for a knowledge representation

a representation should:

- have descriptive and explanatory power

  (allow backtracking into observations)

- be practical and convenient for further application

  (retain feature structure)

- be reasonably true to human performance

  (handle streaming and analogy!)

- handle patchy data

  (provide support for generalisation, defaults and constraints)

- be computationally habitable

  (not grow superlinearly with data input)

- be general

  (not tightly bound to some task)

# requirements for a knowledge representation

a representation should:

- have descriptive and explanatory power

  (allow backtracking into observations)

- be practical and convenient for further application

  (retain feature structure)

- be reasonably true to human performance

  (handle streaming and analogy!)

- handle patchy data

  (provide support for generalisation, defaults and constraints)

- be computationally habitable

  (not grow superlinearly with data input)

- be general

  (not tightly bound to some task)

# the pattern is more interesting than the data points

- new services, e.g. Internet of Things, sensor networks, human-generated data, universal logging: streaming data from many devices, many people, many levels of abstraction

# high-dimensional computing

the approach suggested by us is

- high-dimensional

  to allow a rich representation

- implemented as a vector space

  mathematically well defined and manageable for implementation

- uses random patterns to index observations

  achieves orthogonality for all practical purposes

- useful as a preprocessing encoding for other further models

---

- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

# high-dimensional computing

the approach suggested by us is

- high-dimensional

  to allow a rich representation

- implemented as a vector space

  mathematically well defined and manageable for implementation

- uses random patterns to index observations

  achieves orthogonality for all practical purposes

- useful as a preprocessing encoding for other further models

---

- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

# high-dimensional computing

the approach suggested by us is

- high-dimensional

  to allow a rich representation

- implemented as a vector space

  mathematically well defined and manageable for implementation

- uses random patterns to index observations

  achieves orthogonality for all practical purposes

- useful as a preprocessing encoding for other further models

---

- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

# high-dimensional computing

the approach suggested by us is

- high-dimensional

  to allow a rich representation

- implemented as a vector space

  mathematically well defined and manageable for implementation

- uses random patterns to index observations

  achieves orthogonality for all practical purposes

- useful as a preprocessing encoding for other further models

---

- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

# high-dimensional computing

the approach suggested by us is

- high-dimensional

  to allow a rich representation

- implemented as a vector space

  mathematically well defined and manageable for implementation

- uses random patterns to index observations

  achieves orthogonality for all practical purposes

- useful as a preprocessing encoding for other further models

---

- Pentti Kanerva. 2009. "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation 1:2.

# distributional semantics

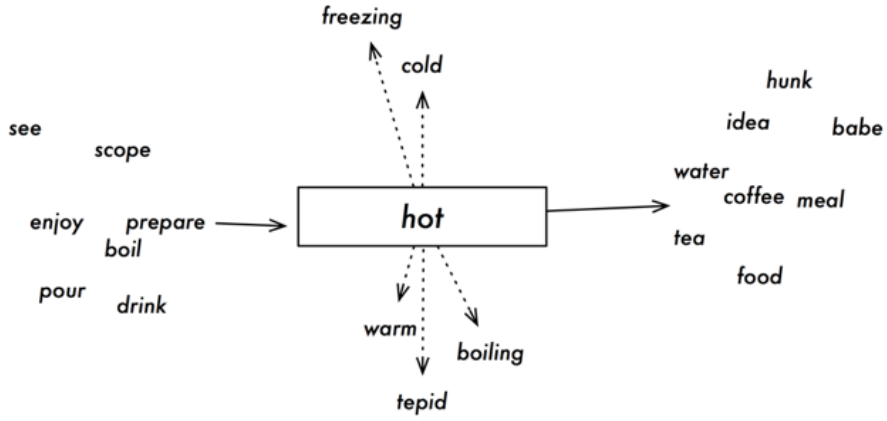
the weather is great in barcelona
the weather is gray in stockholm
the weather is hot in needles
the climate is passable in nice
the weather is chilly in helsinki
the weather is nippy in moscow
the weather is nice in hong kong
the weather in syktyvkar is balmy
the climate is chilly at the office
the 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

the weather is great in barcelona
the weather is gray in stockholm
the weather is hot in needles
the climate is passable in nice
the weath
the weath
the weath
the weath
the clima
the tea i
i drink tea
a hot meal will make you feel better
enjoy your hot beverages




---

- Zellig Harris. 1968. Mathematical structures of language.

# implementational model: semantic vectors

- (relatively) low-dimensional projections of high-dimensional observational data and their cooccurrence
- generalises over cooccurrence events
- quantifies distributional data
- (surprisingly) not (often) systematically used in broader data contexts
- very popular in recent approaches (aka embeddings)
- risk of low explainability and blackboxiness

# implementational model: semantic vectors

- (relatively) low-dimensional projections of high-dimensional observational data and their cooccurrence
- generalises over cooccurrence events
- quantifies distributional data
- (surprisingly) not (often) systematically used in broader data contexts
- very popular in recent approaches (aka embeddings)
- risk of low explainability and blackboxiness

# implementational model: semantic vectors

- (relatively) low-dimensional projections of high-dimensional observational data and their cooccurrence
- generalises over cooccurrence events
- quantifies distributional data
- (surprisingly) not (often) systematically used in broader data contexts
- very popular in recent approaches (aka embeddings)
- risk of low explainability and blackboxiness

# implementational model: semantic vectors

- (relatively) low-dimensional projections of high-dimensional observational data and their cooccurrence
- generalises over cooccurrence events
- quantifies distributional data
- (surprisingly) not (often) systematically used in broader data contexts
- very popular in recent approaches (aka embeddings)
- risk of low explainability and blackboxiness

# implementational model: semantic vectors

- (relatively) low-dimensional projections of high-dimensional observational data and their cooccurrence
- generalises over cooccurrence events
- quantifies distributional data
- (surprisingly) not (often) systematically used in broader data contexts
- very popular in recent approaches (aka embeddings)
- risk of low explainability and blackboxiness

# implementational model: semantic vectors

- (relatively) low-dimensional projections of high-dimensional observational data and their cooccurrence
- generalises over cooccurrence events
- quantifies distributional data
- (surprisingly) not (often) systematically used in broader data contexts
- very popular in recent approaches (aka embeddings)
- risk of low explainability and blackboxiness

# random indexing of human language

- sparse random index vectors or labels for basic features
- aggregated (and thus denser) context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- operations on vectors:
  - addition which yields a vector similar to its inputs
  - permutation which yields a near-orthogonal vector to its input and allows for e.g. sequences or tensors to be represented in one vector
- allows for explicit feature engineering if necessary

# random indexing of human language

- sparse random index vectors or labels for basic features
- aggregated (and thus denser) context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- operations on vectors:
  - addition which yields a vector similar to its inputs
  - permutation which yields a near-orthogonal vector to its input and allows for e.g. sequences or tensors to be represented in one vector
- allows for explicit feature engineering if necessary

# random indexing of human language

- sparse random index vectors or labels for basic features
- aggregated (and thus denser) context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- operations on vectors:
  - addition which yields a vector similar to its inputs
  - permutation which yields a near-orthogonal vector to its input and allows for e.g. sequences or tensors to be represented in one vector
- allows for explicit feature engineering if necessary

# random indexing of human language

- sparse random index vectors or labels for basic features
- aggregated (and thus denser) context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- operations on vectors:
  - addition which yields a vector similar to its inputs
  - permutation which yields a near-orthogonal vector to its input and allows for e.g. sequences or tensors to be represented in one vector
- allows for explicit feature engineering if necessary

# random indexing of human language

- sparse random index vectors or labels for basic features
- aggregated (and thus denser) context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- operations on vectors:
  - addition which yields a vector similar to its inputs
  - permutation which yields a near-orthogonal vector to its input and allows for e.g. sequences or tensors to be represented in one vector
- allows for explicit feature engineering if necessary

# random indexing of human language

- sparse random index vectors or labels for basic features
- aggregated (and thus denser) context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- operations on vectors:
  - addition which yields a vector similar to its inputs
  - permutation which yields a near-orthogonal vector to its input and allows for e.g. sequences or tensors to be represented in one vector
- allows for explicit feature engineering if necessary

# random indexing of human language

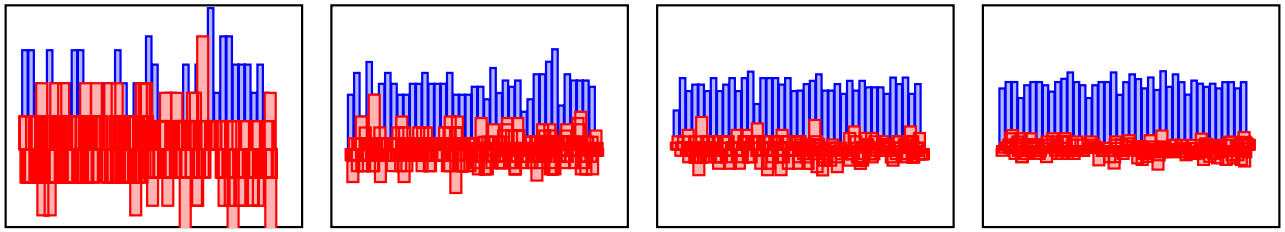
- sparse random index vectors or labels for basic features
- aggregated (and thus denser) context vectors for cooccurrences of features
- similarity between vectors can be measured by cosine
- operations on vectors:
  - addition which yields a vector similar to its inputs
  - permutation which yields a near-orthogonal vector to its input and allows for e.g. sequences or tensors to be represented in one vector
- allows for explicit feature engineering if necessary

# quantitative characteristics

1. random patterns instead of isolated dimensions: allows a vastly more features in a space
2. permutations allow several aggregated contexts simultaneously

# quantitative characteristics

## how much will the randomness cost in retrievability?



generate random vector $\vec{r}$ and $\vec{g}$ with near zero cosine; add twenty other random vectors to $\vec{r}$ to make $\vec{r_s}$; measure cosine between $\vec{r}$ and $\vec{r_s}$ and the cosine between $\vec{g}$ and $\vec{r_s}$; blue bars give $\vec{r}$ cosines and red bars give $\vec{g}$ cosines for twenty items in 100, 500, 1000, 2000 dimensions

```
   hot  [0, 0, 0, 0, 0, 0]
chilly  [0, 0, 0, 0, 0, 0]
```

---

- Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random Indexing of Text Samples for Latent Semantic Analysis. CogSci.
- Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a Means to Encode Order in Word Space. CogSci.

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

weather: [ 0, -1, 0, 1, 0, 0]

hot [0, -1, 0, 1, 0, 0]
chilly [0, -1, 0, 1, 0, 0]

cosine: 1.0

---

- Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random Indexing of Text Samples for Latent Semantic Analysis. CogSci.
- Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a Means to Encode Order in Word Space. CogSci.

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

```
weather:    [ 0, -1, 0, 1, 0, 0]
climate:    [ 1, 0, -1, 0, 0, 0]
```

```
   hot  [1, -1, -1, 1, 0, 0]
chilly  [1, -1, -1, 1, 0, 0]
```

cosine: 1.0

---


- Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random Indexing of Text Samples for Latent Semantic Analysis. CogSci.
- Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a Means to Encode Order in Word Space. CogSci.

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

the weather turned hot ... the climate seems chilly ...

```
  weather:    [ 0, -1, 0, 1, 0, 0]
  climate:    [ 1, 0, -1, 0, 0, 0]
   turned:    [ 0, 0, 0, 0, -1, 1]
    seems:    [ -1, 0, 0, 0, 1, 0]
```

```
    hot  [2, -1, -2, 1, -1, 1]
chilly  [0, -2, -1, 2, 1, 0]
```

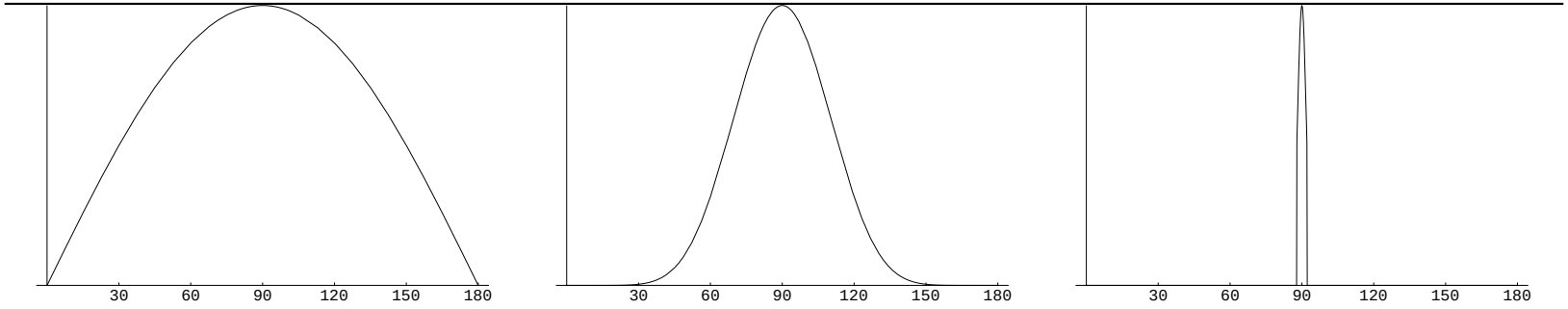
cosine:  0.4564

---

- Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random Indexing of Text Samples for Latent Semantic Analysis. CogSci.
- Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a Means to Encode Order in Word Space. CogSci.

# high-dimensional space

probability distribution of the angle between two randomly chosen points in 3-, 10-, and 1000-dimensional space




in high dimensional space the transitivity of distance goes awry, and the notion of a cluster or a centroid is less useful

---

- Jussi Karlgren, Anders Holst, and Magnus Sahlgren. 2008. Filaments of Meaning in Word Space. ECIR.

# plenty of parameters: the craft of distributional models

distributional models come in many varieties and all need to handle a number of challenges related to the nature of human language and to processing sparse data

- **context** : what is the relevant semantic similarity?
- **direction** : are the left and the right contexts different?
- **weighting** : how do we distinguish unusual and trivial items and collocations from relevant and interesting ones?
- **streaming** : can we learn continuously or do we need to compile?
- **more than single words** : do we need to handle multi-word terms and collocations?
- **quality assurance** : are we doing the right thing?
- **the black hole of semantics** : with large amounts of data, everything risks being like everything else

# plenty of parameters: the craft of distributional models

distributional models come in many varieties and all need to handle a number of challenges related to the nature of human language and to processing sparse data

context : what is the relevant semantic similarity?

direction : are the left and the right contexts different?

weighting : how do we distinguish unusual and trivial items and collocations from relevant and interesting ones?

streaming : can we learn continuously or do we need to compile?

more than single words : do we need to handle multi-word terms and collocations?

quality assurance : are we doing the right thing?

the black hole of semantics : with large amounts of data, everything risks being like everything else

# plenty of parameters: the craft of distributional models

distributional models come in many varieties and all need to handle a number of challenges related to the nature of human language and to processing sparse data

context : what is the relevant semantic similarity?

direction : are the left and the right contexts different?

weighting : how do we distinguish unusual and trivial items and collocations from relevant and interesting ones?

streaming : can we learn continuously or do we need to compile?

more than single words : do we need to handle multi-word terms and collocations?

quality assurance : are we doing the right thing?

the black hole of semantics : with large amounts of data, everything risks being like everything else

# plenty of parameters: the craft of distributional models

distributional models come in many varieties and all need to handle a number of challenges related to the nature of human language and to processing sparse data

context : what is the relevant semantic similarity?

direction : are the left and the right contexts different?

weighting : how do we distinguish unusual and trivial items and collocations from relevant and interesting ones?

streaming : can we learn continuously or do we need to compile?

more than single words : do we need to handle multi-word terms and collocations?

quality assurance : are we doing the right thing?

the black hole of semantics : with large amounts of data, everything risks being like everything else

# plenty of parameters: the craft of distributional models

distributional models come in many varieties and all need to handle a number of challenges related to the nature of human language and to processing sparse data

context : what is the relevant semantic similarity?

direction : are the left and the right contexts different?

weighting : how do we distinguish unusual and trivial items and collocations from relevant and interesting ones?

streaming : can we learn continuously or do we need to compile?

more than single words : do we need to handle multi-word terms and collocations?

quality assurance : are we doing the right thing?

the black hole of semantics : with large amounts of data, everything risks being like everything else

# plenty of parameters: the craft of distributional models

distributional models come in many varieties and all need to handle a number of challenges related to the nature of human language and to processing sparse data

context : what is the relevant semantic similarity?

direction : are the left and the right contexts different?

weighting : how do we distinguish unusual and trivial items and collocations from relevant and interesting ones?

streaming : can we learn continuously or do we need to compile?

more than single words : do we need to handle multi-word terms and collocations?

quality assurance : are we doing the right thing?

the black hole of semantics : with large amounts of data, everything risks being like everything else

# plenty of parameters: the craft of distributional models

distributional models come in many varieties and all need to handle a number of challenges related to the nature of human language and to processing sparse data

context : what is the relevant semantic similarity?

direction : are the left and the right contexts different?

weighting : how do we distinguish unusual and trivial items and collocations from relevant and interesting ones?

streaming : can we learn continuously or do we need to compile?

more than single words : do we need to handle multi-word terms and collocations?

quality assurance : are we doing the right thing?

the black hole of semantics : with large amounts of data, everything risks being like everything else

# plenty of parameters: the craft of distributional models

distributional models come in many varieties and all need to handle a number of challenges related to the nature of human language and to processing sparse data

**context** : what is the relevant semantic similarity?

**direction** : are the left and the right contexts different?

**weighting** : how do we distinguish unusual and trivial items and collocations from relevant and interesting ones?

**streaming** : can we learn continuously or do we need to compile?

**more than single words** : do we need to handle multi-word terms and collocations?

**quality assurance** : are we doing the right thing?

**the black hole of semantics** : with large amounts of data, everything risks being like everything else

loads of sentences, represented as sums of features (2000 dimensions; features are 10 non-zero cells):

loads of sentences, represented as sums of features (2000 dimensions; features are 10 non-zero cells):

(a) words

loads of sentences, represented as sums of features (2000 dimensions; features are 10 non-zero cells):

(a) words

(b) constructional elements: negations, amplifiers, ...

loads of sentences, represented as sums of features (2000 dimensions; features are 10 non-zero cells):

(a) words

(b) constructional elements: negations, amplifiers, ...

find neighbours to:

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

loads of sentences, represented as sums of features (2000 dimensions; features are 10 non-zero cells):

(a) words

(b) constructional elements: negations, amplifiers, ...

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 (2000 dimensions; features are 10 non-zero cells):

(a) words

(b) constructional elements: negations, amplifiers, ...

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 (2000 dimensions; features are 10 non-zero cells):

(a) words

(b) constructional elements: negations, amplifiers, ...

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.

in the same representational space. great for hypothesis testing, and for differing tasks!

# Gavagai tools

- text clustering: term suggestion, concept definition explorer.gavagai.se
- media monitoring: term suggestion monitor.gavagai.se
- living lexicon in 45 languages: lexicon.gavagai.se
- all available through APIs

# Gavagai tools

- text clustering: term suggestion, concept definition explorer.gavagai.se
- media monitoring: term suggestion monitor.gavagai.se
- living lexicon in 45 languages: lexicon.gavagai.se
- all available through APIs

# Gavagai tools

- text clustering: term suggestion, concept definition explorer.gavagai.se
- media monitoring: term suggestion monitor.gavagai.se
- living lexicon in 45 languages: lexicon.gavagai.se
- all available through APIs

# Gavagai tools

- text clustering: term suggestion, concept definition explorer.gavagai.se
- media monitoring: term suggestion monitor.gavagai.se
- living lexicon in 45 languages: lexicon.gavagai.se
- all available through APIs

# why not just use existing models?

- we believe this to be the only sustainable computational approach for knowledge representation in applications where data is streaming, varied, and at real-world scale
- at gavagai we have found no reason to depart from the basic premises for the purpose of processing human language
- this is in contrast to
  - localist representations,
  - compiling models,
  - and end-to-end black boxes.

# why not just use existing models?

- we believe this to be the only sustainable computational approach for knowledge representation in applications where data is streaming, varied, and at real-world scale
- at gavagai we have found no reason to depart from the basic premises for the purpose of processing human language
- this is in contrast to
  - localist representations,
  - compiling models,
  - and end-to-end black boxes.

# why not just use existing models?

- we believe this to be the only sustainable computational approach for knowledge representation in applications where data is streaming, varied, and at real-world scale
- at gavagai we have found no reason to depart from the basic premises for the purpose of processing human language
- this is in contrast to
  - localist representations,
  - compiling models,
  - and end-to-end black boxes.

# why not just use existing models?

- we believe this to be the only sustainable computational approach for knowledge representation in applications where data is streaming, varied, and at real-world scale
- at gavagai we have found no reason to depart from the basic premises for the purpose of processing human language
- this is in contrast to
  - localist representations,
  - compiling models,
  - and end-to-end black boxes.

# why not just use existing models?

- we believe this to be the only sustainable computational approach for knowledge representation in applications where data is streaming, varied, and at real-world scale
- at gavagai we have found no reason to depart from the basic premises for the purpose of processing human language
- this is in contrast to
  - localist representations,
  - compiling models,
  - and end-to-end black boxes.

# why not just use existing models?

- we believe this to be the only sustainable computational approach for knowledge representation in applications where data is streaming, varied, and at real-world scale
- at gavagai we have found no reason to depart from the basic premises for the purpose of processing human language
- this is in contrast to
  - localist representations,
  - compiling models,
  - and end-to-end black boxes.

# take home

1. 
   - handing hypotheses and providing explanatory power are important ...
   - ... as is computational habitality ...
   - ... neither is optional

# take home

1. 
   - handing hypotheses and providing explanatory power are important ...
   - ... as is computational habitality ...
   - ... neither is optional
2. 
   - feature engineering is a useful method to understand the world ...
   - ... knowledge representations for processing large amounts of data should support it

# take home

1. 
   - handing hypotheses and providing explanatory power are important ...
   - ... as is computational habitality ...
   - ... neither is optional
2. 
   - feature engineering is a useful method to understand the world ...
   - ... knowledge representations for processing large amounts of data should support it
3. 
   - neurophysiological plausibility is optional ...
   - ... as is symbolic explicitness and lucidity

# take home

1. 
   - handing hypotheses and providing explanatory power are important ...
   - ... as is computational habitality ...
   - ... neither is optional
2. 
   - feature engineering is a useful method to understand the world ...
   - ... knowledge representations for processing large amounts of data should support it
3. 
   - neurophysiological plausibility is optional ...
   - ... as is symbolic explicitness and lucidity
4. 
   - we have APIs!