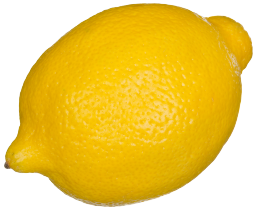
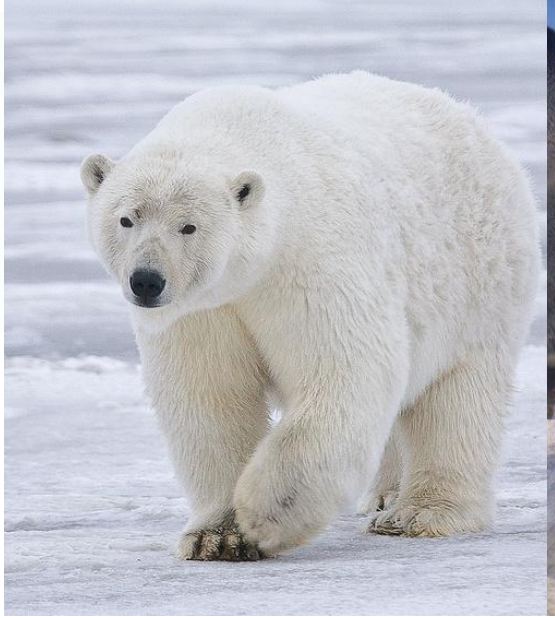
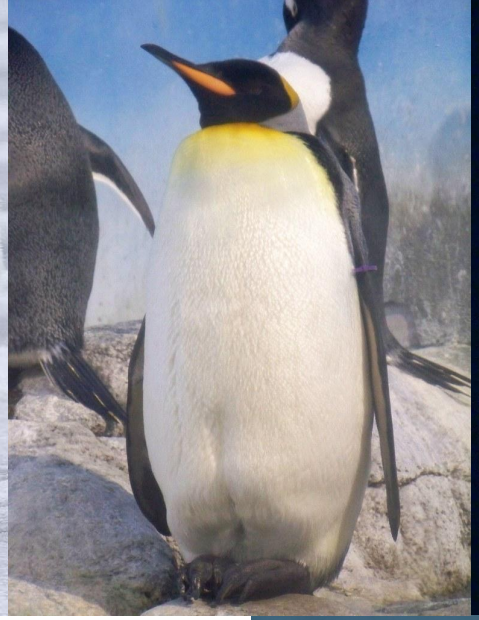
Can patterns of word usage tell us what *lemon* and *moon* have in common?

Analyzing the semantic content of distributional semantic models







What makes word meanings similar?

What are we going to do?

- Why are distributional semantic models interesting?
- Distributional semantics & distributional semantic models
- What we know & what we don’t know

  (and how we might find out)
- Hands-on practice (if time)

Introduction

What is word meaning?

What is word meaning?

What do distributional representations of word meaning contain?

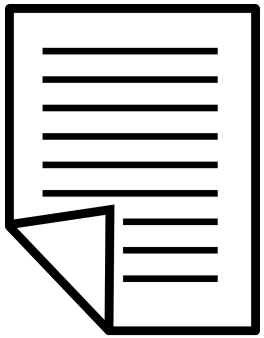
How can we represent the meaning of words?

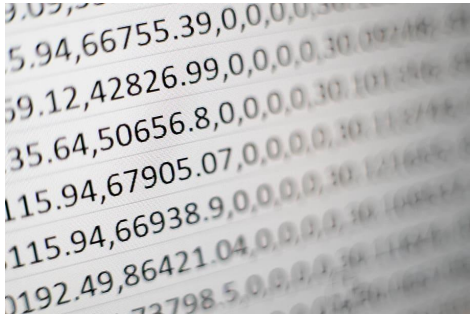
hand-crafted	Dictionary definitions
human-elicited	Feature vectors
text	Distributional vectors

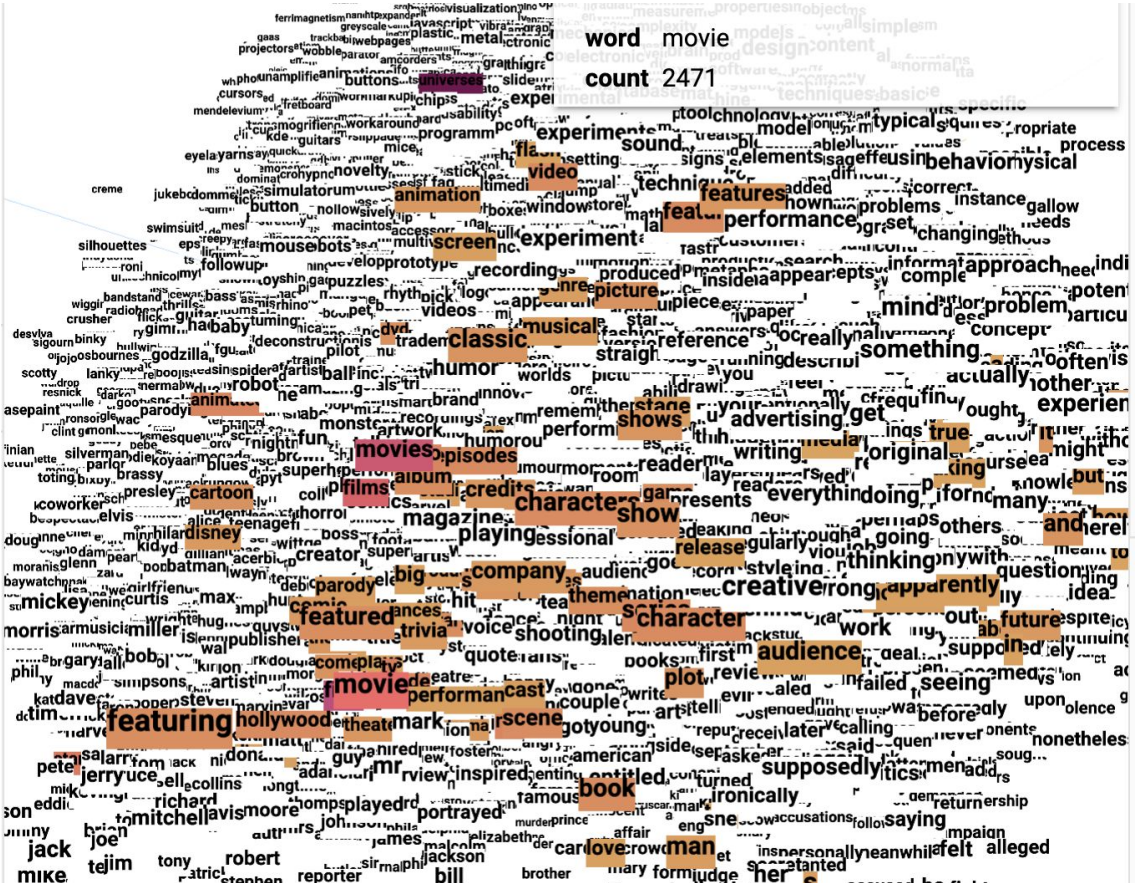
Who can read these representations?

Human	Dictionary definitions
Human & computer	Hand-crafted computational lexica Human-elicited feature-vectors
Computer	Distributional vectors

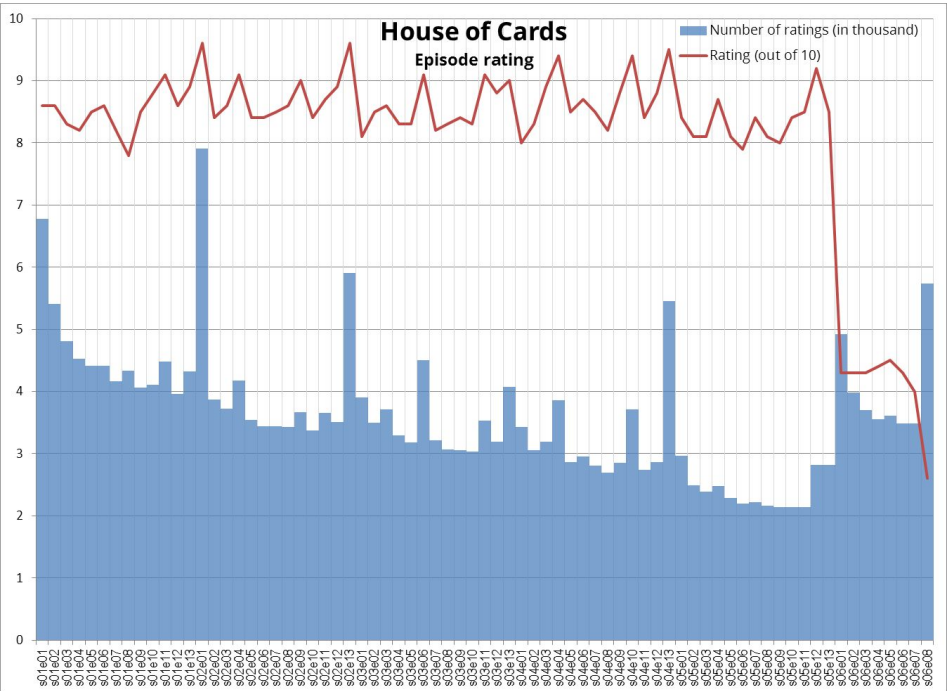
Meaning from massive amouts of text

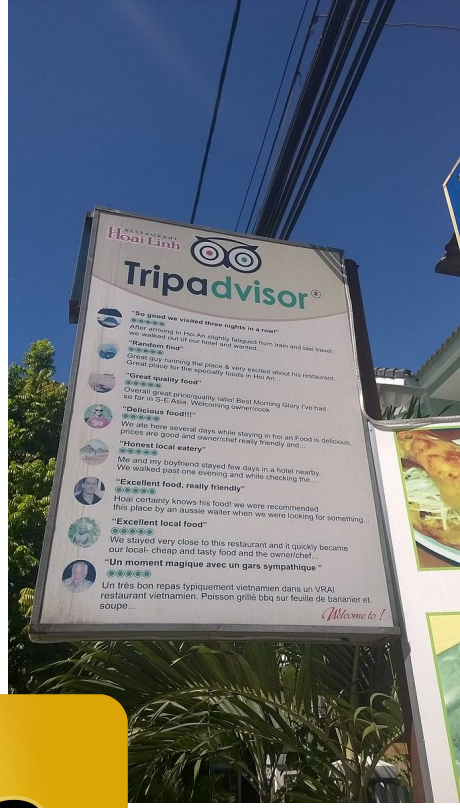






NLP applications need word meaning





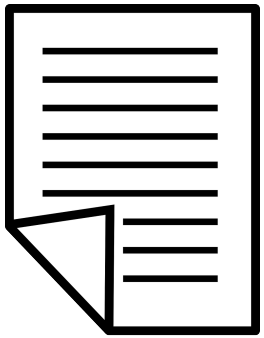
IMDb

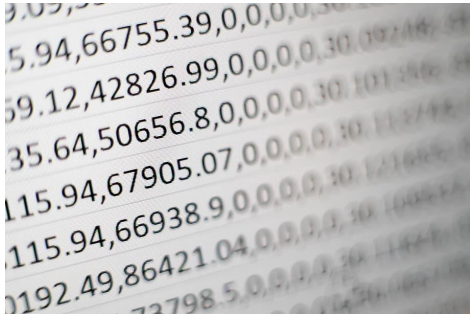
Example: Sentiment mining

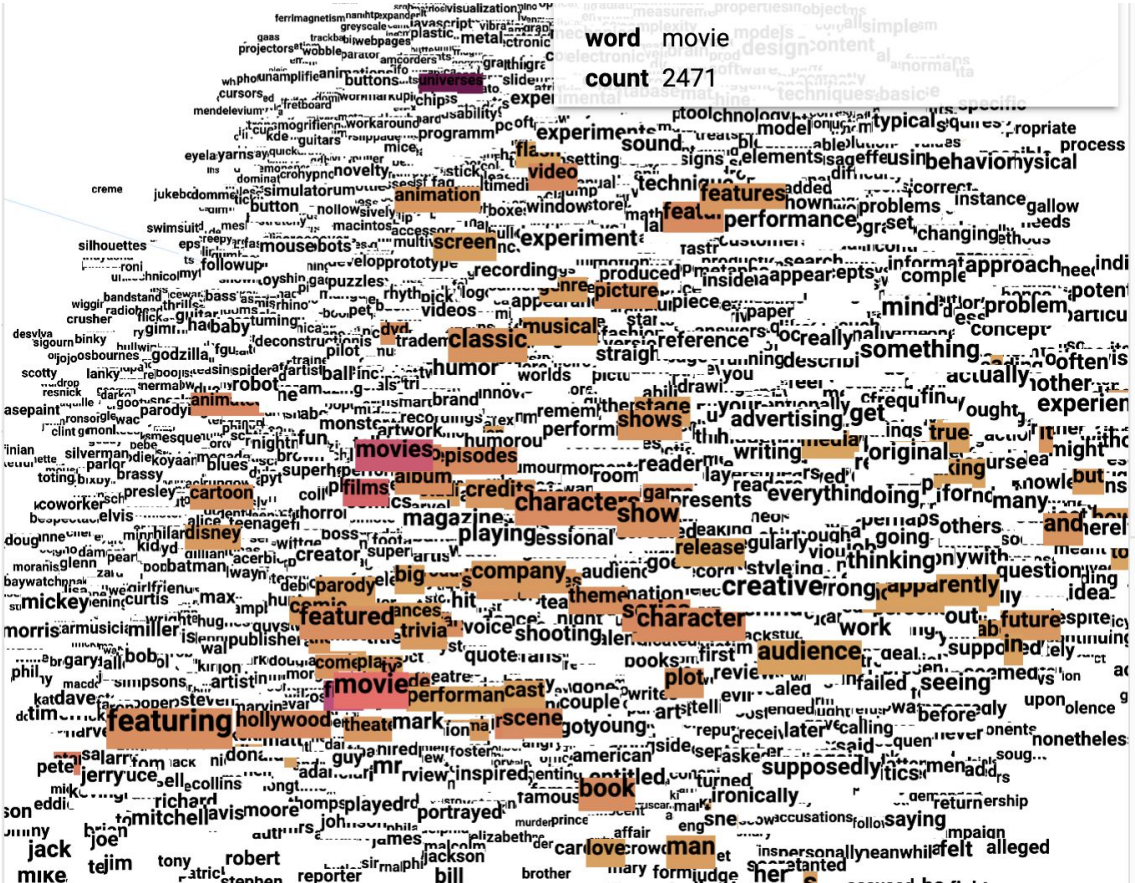
This movie is a gosh darn masterpiece. It will make you belly laugh, it will chill you to the bone, and it will make you shed a tear. This movie will stay with you long after the credits are over.

How has this film has won the best film Oscar? What a travesty. It's inane, rebarbative, empty, childish, overlong, silly, dull, cluttered and generally irritating. The social message, presuming there is one, is garbled, superficial and hollow.

Meaning vectors from massive amounts of text







Embeddings - the good

- Wide coverage
- Compatible with ML and DL → boost
- Generalization
- No hand-crafting
- Meaning based on actual usage

Embeddings - the bad

- Not transparent
- No reasoning
- Evaluation is not perfect
- Usually require a lot of data

Embeddings - the ugly

- Bias

direct stakeholders. For example, Speer (2017) found that a sentiment analysis system rated reviews of Mexican restaurants as more negative than other types of food with similar star ratings, because of associations between the word *Mexican* and words with negative sentiment in the larger corpus on which the word embeddings were trained. (See also Kiritchenko and Mohammad, 2018.) In these and other ways, pre-existing bi-

Bender & Friedman 2018

Therefore:

What aspects of semantics are represented by embedding vectors?

A bit of background: Distributional semantics

Meaning = Use

“You shall know a word by the company it keeps”

(Firth, J. R. 1957:11)

What do we know about a word **X**?

(1) The **X** wasn't very hot though , made in a filter pot , but it was good .
(2) She sees that there is a cup of steaming hot **X** awaiting him and the two chat informally as she presents the rules of the center and explains procedures .
(3) Eugene put a spoonful of powdered **X** into his cup and then filled it with hot water .
(4) Would you like a drink , or **X** " ? ?
(5) She could not face **X** or tea without milk , and was always craving types of food that were not available aboard a sailing ship .

(taken from the Brown corpus)

What do we know about a word **X**?

- supposed to be hot
- comes in a cup
- it is poured
- people drink it
- people offer it
- there are shops for if
- comes in powder
- similar to tea
- people add milk to it

What do we know about a word X?

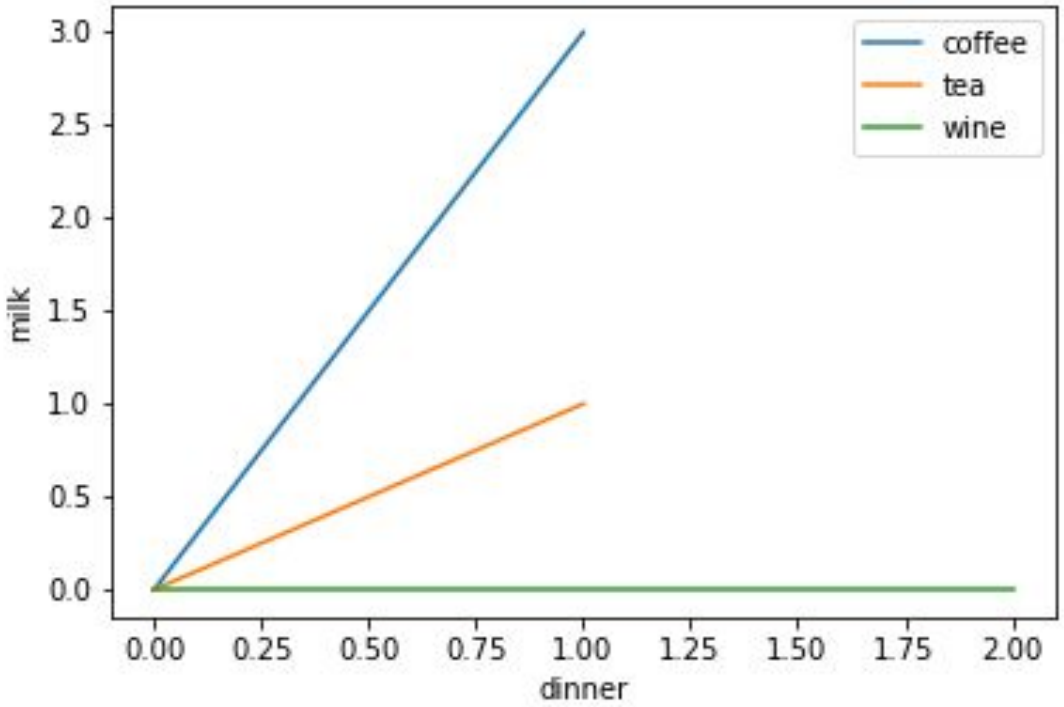
- supposed to be hot
- comes in a cup
- it is poured
- people drink it
- people offer it
- there are shops for if
- comes in powder
- similar to tea
- people add milk to it

Distributional semantic models

A simple co-occurrence matrix

	coffee	tea	milk	dinner	wine
coffee	3	1	3	1	0
tea	1	1	1	1	0
milk	2	1	4	0	0
dinner	1	1	0	1	2
wine	0	0	0	2	8

Vectors


coffee
tea
wine
milk
dinner
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

Defining context

What counts as ‘context’?

Defining context

What counts as ‘context’?

- Sentence
- Number of words around the target word (e.g. +-4)
- Syntactically defined (arguments)

Defining context: consequences

Nearest neighbors of *dog*

(Baroni & Boleda)

2-word window

- cat
- horse
- fox
- pet
- rabbit
- pig
- animal
- mongrel
- sheep
- pigeon

30-word window

- kennel
- puppy
- pet
- bitch
- terrier
- rottweiler
- canine
- cat
- to bark
- Alsatian

Creating models

Linguistic tradition:
count co-occurrences

Computer science tradition:
Machine Learning inspired by language models

→ Similar results

New NLP model: BERT

***Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors**

Marco Baroni and **Georgiana Dinu** and **Germán Kruszewski**
Center for Mind/Brain Sciences (University of Trento, Italy)
(marco.baroni|georgiana.dinu|german.kruszewski)@unitn.it

A Primer in BERTology: What we know about how BERT works

Anna Rogers, Olga Kovaleva, Anna Rumshisky
Department of Computer Science, University of Massachusetts Lowell
Lowell, MA 01854
{arogers, okovalev, arum}@cs.uml.edu

Counting

More sophisticated counting:

- Positive pointwise mutual information score
  ~How likely are two words to co-occur as opposed to occurring separately?
- PPMI + SVD (singular value decomposition)

(e.g. Levi et al. 2015)

Strudel

- POS patterns to extract property mentions
- Semantic links between properties and concepts

(Baroni et al. 2010)

Neural Language models: an intuition

Task: guess the next word (Bengio 2003)

You will notice it if I say an unexpected ___.

Task: distinguish correct from incorrect word sequence
(Collobert & Weston 2008)

You will notice it if I say an unexpected word.
vs
You will notice it if I say an unexpected strawberry.

Machine Learning: Word2vec

- Task (variation 1): predict word given context
- Task (variation 2): predict context given word
- Remove intermediate layers → more efficient

(Mikolov 2013)

Big model available for download (GoogleNews) (~100 billion words)

Shown to improve NLP tasks
(e.g. Zhou 2015, Socher et al. 2013)

Easy to use
(Gensim package for python)

Machine Learning: BERT

(1) Create embeddings by means of (1) masked token prediction and (2) sentence prediction
(2) Task-specific fine-tuning

Architectures: stacked transformer encoder layers

Contextualized representations (1 representation per token in the corpus)

Analyzing embeddings: what we know

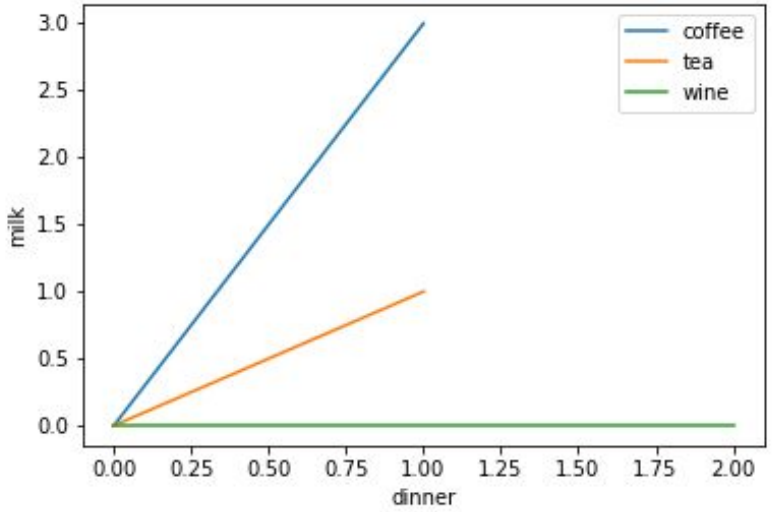
What’s a good vector?

Evaluation

In general: Compare model output to human judgments

Semantic space:

- No definitions (you cannot ‘read’ a list of numbers)
- A word is defined by its relation to other words
- Relation: distance



$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Semantic similarity and relatedness judgements

→ Humans rank word pairs according to similarity and relatedness

→ Average over multiple annotators

→ Rankings produced by the distributional model should correlate with the human judgements (Spearman Rho correlation)

Similarity in the distributional model is expressed as the cosine between two vectors

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Evaluation: Semantic similarity and relatedness

	Similar	**Related**
coffee-tea	yes	yes
coffee-cup	no	yes
gasoline-cup	no	yes

Evaluation: Analogical reasoning

Are individual semantic properties are encoded in (patterns of) dimensions?

Man:woman ≈ king:queen *(Mikolov et al. 2013)*

King - man + woman ≈ queen

0	0	1	1	female
1	1	0	0	male
1	0	0	1	royal

Heavily criticized
(e.g. Linzen 2016)

Analogy task

given	given	given	predict
bangkok	thailand	paris	france
bangkok	thailand	rome	italy
bangkok	thailand	stockholm	sweden
bangkok	thailand	tehran	iran
bangkok	thailand	tokyo	japan
beijing	china	berlin	germany
beijing	china	bern	switzerland

Analogy task criticism

given	given	given	predict
bangkok	thailand	paris	france
bangkok	thailand	rome	italy
bangkok	thailand	stockholm	sweden
bangkok	thailand	tehran	iran
bangkok	thailand	tokyo	japan
beijing	china	berlin	germany
beijing	china	bern	switzerland

Excluded from possible answers

Which models perform best?

Count, predict, which corpus?

Intrinsic evaluation

Extensive comparisons using PPMI, PPMI-SVD, Word2vec models (trained on Wikipedia) lead to contradictory results (Baroni et al. 2014, Levy et a. 2015).

Results vary, hyperparameters have an impact, scores are sometimes very close

What does evaluation tell us?

“Performance on **downstream tasks** is **not consistent across tasks**, and m**ay not be consistent with intrinsic evaluations**. Comparing performance across tasks may provide insight into the information encoded by an embedding, but **we should not expect any specific task to act as a proxy for abstract quality**. “

From: Evaluation methods for unsupervised word embeddings (Schnabel et al. 2015, p. 304)

BERT: no intrinsic evaluation, only task specific

What is actually ‘in’ a vector?

Vectors for semantic reasoning

King =

royal	male	adult	person

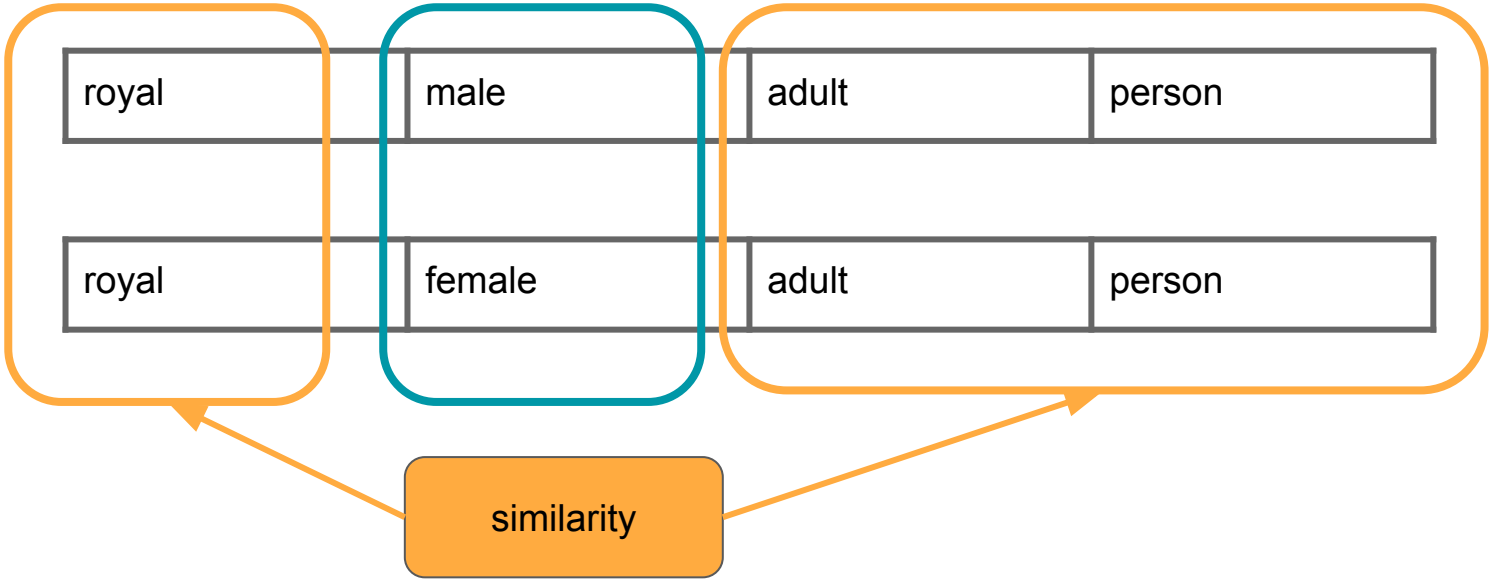
Queen =

royal	female	adult	person

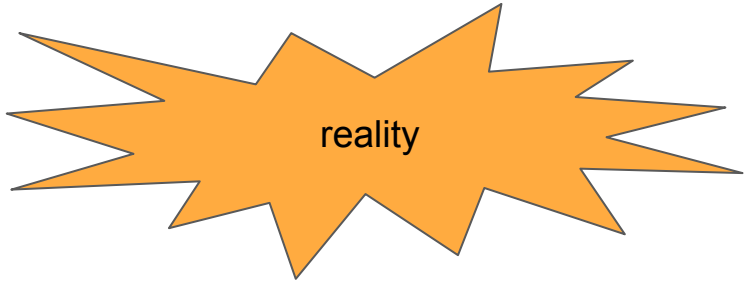
Vectors for semantic reasoning

King =

Queen =



Vectors for semantic reasoning?



king =	-0.45645	9.13536	-2.34572	10.23508	1.235645	2.236657	[...]
queen =	-1.23462	5.23765	-0.84562	3.146494	11.23535	-4.12453	[...]

Similarity?

Cosine

Towards vectors for semantic reasoning

king =	royal	male	adult	person
queen =	royal	female	adult	person

king =	-0.45645	9.13536	-2.34572	10.23508	1.235645	2.236657	[...]
queen =	-1.23462	5.23765	-0.84562	3.146494	11.23535	-4.12453	[...]

From embeddings to semantic features

Learn transformation from embedding space to human-elicited feature space
(Fagharasan 2015, Herbelot 2015, Derby et al. 2018)

- Overall correlation scores
- Improved performance
- Problematic datasets

No insights about underlying mechanisms and model potential

We want to know more!

Analysis in terms of semantic properties

- Knowledge vs what is mentioned in text
- How is knowledge expressed in text?
- Model sensitivity to linguistic evidence
- Potential of analysis methods

What can we expect from linguistic context?

Linguistic hypotheses

- Gricean Maxims
- Typicality
- Variability
- Affordedness

What can we expect from linguistic context?

Linguistic hypotheses

- Impliedness (Gricean Maxims)
- Typicality
- Variability
- Affordedness

Highly implied knowledge is not made explicit

→ Violation of the Gricean maxim of quantity

What can we expect from linguistic context?

Linguistic hypotheses

- Impliedness (Gricean Maxims)
- Typicality
- Variability
- Affordedness

Mentioned:
Concepts illustrating properties
(Veale 2011, Veale & Hao 2007, Veale 2013)

As white as snow, as red as blood, as black as ebony wood

Not mentioned:
Properties typical of a concept

Green broccoli

What can we expect from linguistic context?

Linguistic hypotheses

- Impliedness (Gricean Maxims)
- Typicality
- Variability
- Affordedness

Mentioned:
Instances of concepts vary with respect to the property:

red/green/yellow bell peppers
brown/black/grey bears

What can we expect from linguistic context?

Linguistic hypotheses

- Impliedness (Gricean Maxims)
- Typicality
- Variability
- Affordedness

Mentioned:
Things instances of concepts usually do/activities they are involved in
Burning candle

Not mentioned:
Possible but unusual activities
Burning candle

(Gibson 1954, Glenberg 1997, Glenberg 2000, Fulda et al. 2017)

Model analysis methods

E.g. Diagnostic classification

- Binary classifier
- Can examples be separated based on their embeddings?

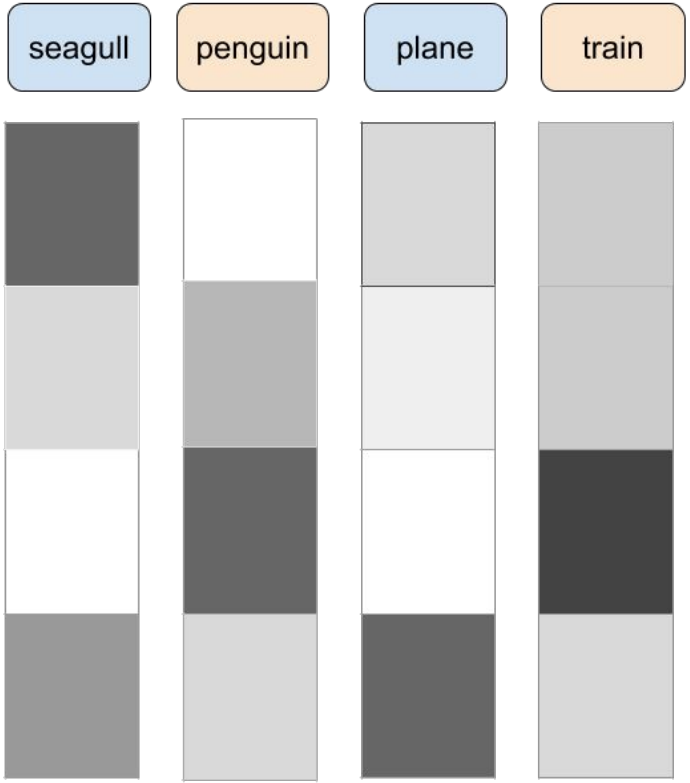
(Hupkes 2018, Belinkov & Glass 2019)

Model analysis methods

E.g. Diagnostic classification

- Binary classifier
- Can examples be separated based on their embeddings?

(Hupkes 2018, Belinkov & Glass 2019)

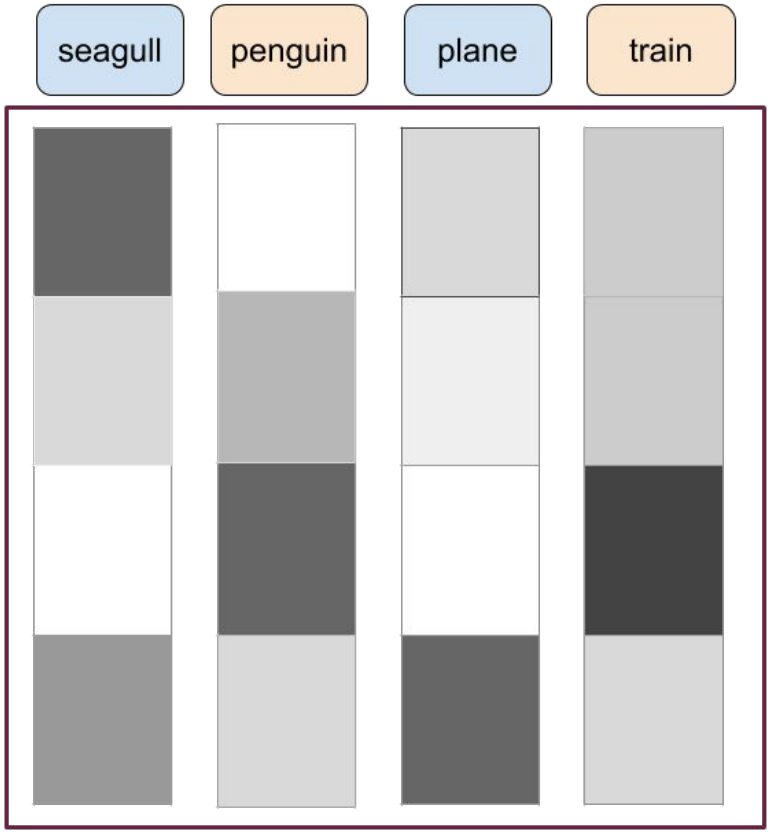


Model analysis methods

E.g. Diagnostic classification

- Binary classifier
- Can examples be separated based on their embeddings?

(Hupkes 2018, Belinkov & Glass 2019)

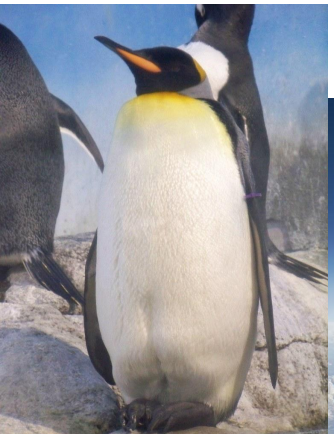


Model analysis methods

Diagnostic classification

- Binary classifier
- Can examples be separated based on their embeddings?

(Hupkes 2018, Belinkov & Glass 2019)

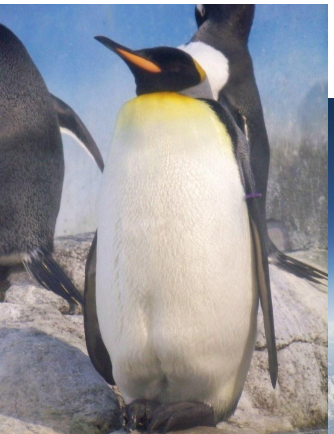


Model analysis methods

Diagnostic classification

- Binary classifier
- Can examples be separated based on their embeddings?

(Hupkes 2018, Belinkov & Glass 2019)

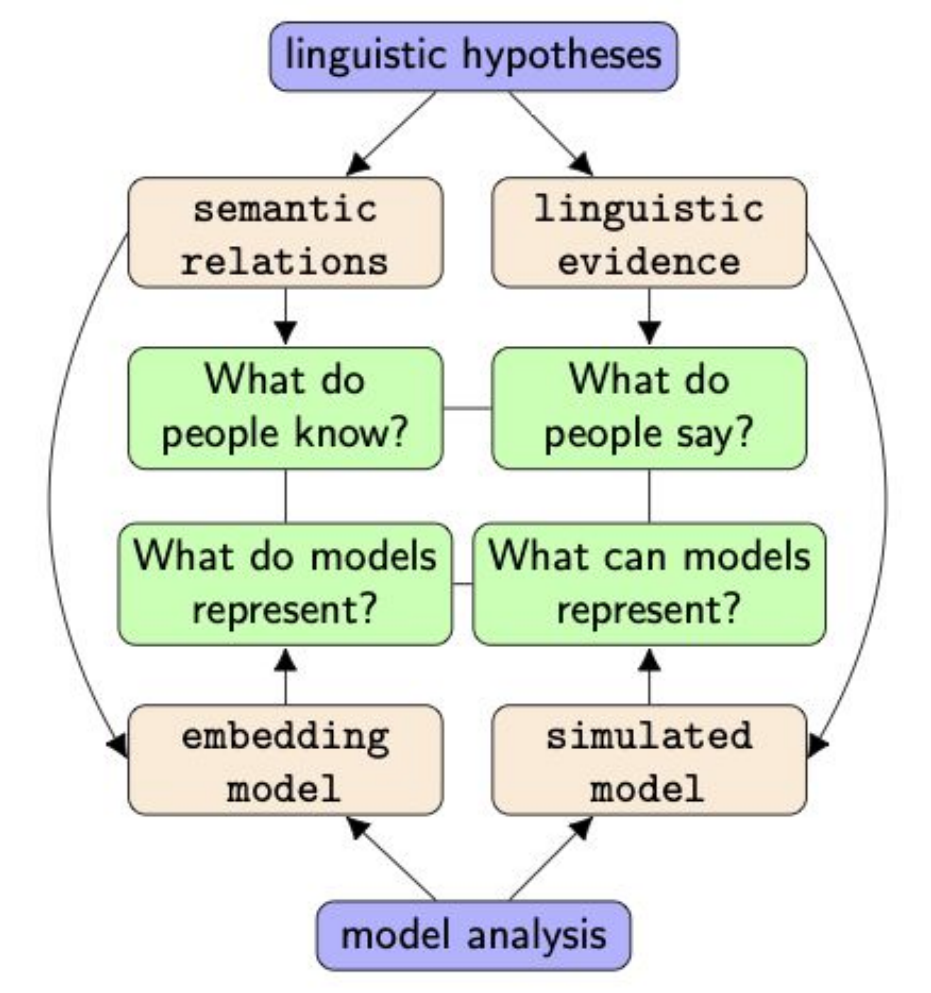


Dataset

- Annotate property-concept pairs with semantic relations
- Distribution of examples:
  - Negative examples similar to positive examples
  - Examples of different semantic categories

[Example]

Plan


linguistic hypotheses
semantic relations
linguistic evidence
What do people know?
What do people say?
What do models represent?
What can models represent?
embedding model
simulated model
model analysis

Summary & conclusions

- Representing word meaning based on context
- Evaluation
- Can we look ‘into’ the representations?
- What can we expect based on linguistic research?
- Model analysis methods

Thank you!

pia.sommerauer@vu.nl

Hands-on practice

References

Baroni & Boleda. Distributional Semantic Models https://www.cs.utexas.edu/~mooney/cs388/slides/dist-sem-intro-NLP-classUT.pdf

Bender, E.M. and Friedman, B., 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, pp.587-604.