Topics in Natural Language Processing

Amaru Cuba Gyllensten

Introduction

Who am I?

- Amaru Cuba Gyllensten
- Industrial PhD Candidate and Researcher at RISE AI.
- Specializes in Language Technology and Distributional Semantic Models.

email me!

Outline

- What and why is Natural Language Processing
- The history of Distributional Semantic Models
- Contemporary language modelling
- Uses of Natural Language Processing

Natural Language Processing

What is it?

The processing, modelling, and exploitation of natural language data.

- Speech recognition

What is it?

The processing, modelling, and exploitation of natural language data.

- Speech recognition
- Machine translation

What is it?

The processing, modelling, and exploitation of natural language data.

- Speech recognition
- Machine translation
- Word segmentation

What is it?

The processing, modelling, and exploitation of natural language data.

- Speech recognition
- Machine translation
- Word segmentation
- Optical Character Recognition

What is it?

The processing, modelling, and exploitation of natural language data.

- Speech recognition
- Machine translation
- Word segmentation
- Optical Character Recognition
- **Natural language understanding**

Applications of Textual understanding

- Sentiment analysis

Applications of Textual understanding

- Sentiment analysis
- Document classification

Applications of Textual understanding

- Sentiment analysis
- Document classification
- Information retrieval

Applications of Textual understanding

- Sentiment analysis
- Document classification
- Information retrieval
- Question answering

Applications of Textual understanding

- Sentiment analysis
- Document classification
- Information retrieval
- Question answering
- Summarization

Text as a data source

Language is a **very** rich and informative data source. Text is abundant.

- 456 000 tweets per minute

Text as a data source

Language is a **very** rich and informative data source. Text is abundant.

- 456 000 tweets per minute
- 510 000 Facebook comments per minute

Text as a data source

Language is a **very** rich and informative data source. Text is abundant.

- 456 000 tweets per minute
- 510 000 Facebook comments per minute
- 293 000 Facebook status updates

Text as a data source

Language is a **very** rich and informative data source. Text is abundant.

- 456 000 tweets per minute
- 510 000 Facebook comments per minute
- 293 000 Facebook status updates
- 16 000 000 text messages

Text as a data source

Language is a **very** rich and informative data source. Text is abundant.

- 456 000 tweets per minute
- 510 000 Facebook comments per minute
- 293 000 Facebook status updates
- 16 000 000 text messages
- 156 000 000 emails

Text as a data source

Language is a **very** rich and informative data source. Text is abundant.

- 456 000 tweets per minute
- 510 000 Facebook comments per minute
- 293 000 Facebook status updates
- 16 000 000 text messages
- 156 000 000 emails
- 103 000 000 spam emails

The richness of text data: The sentiment neuron

Performant unsupervised language models

$\Downarrow$

cheap applications leveraging language **understanding**.

The richness of text data: The sentiment neuron

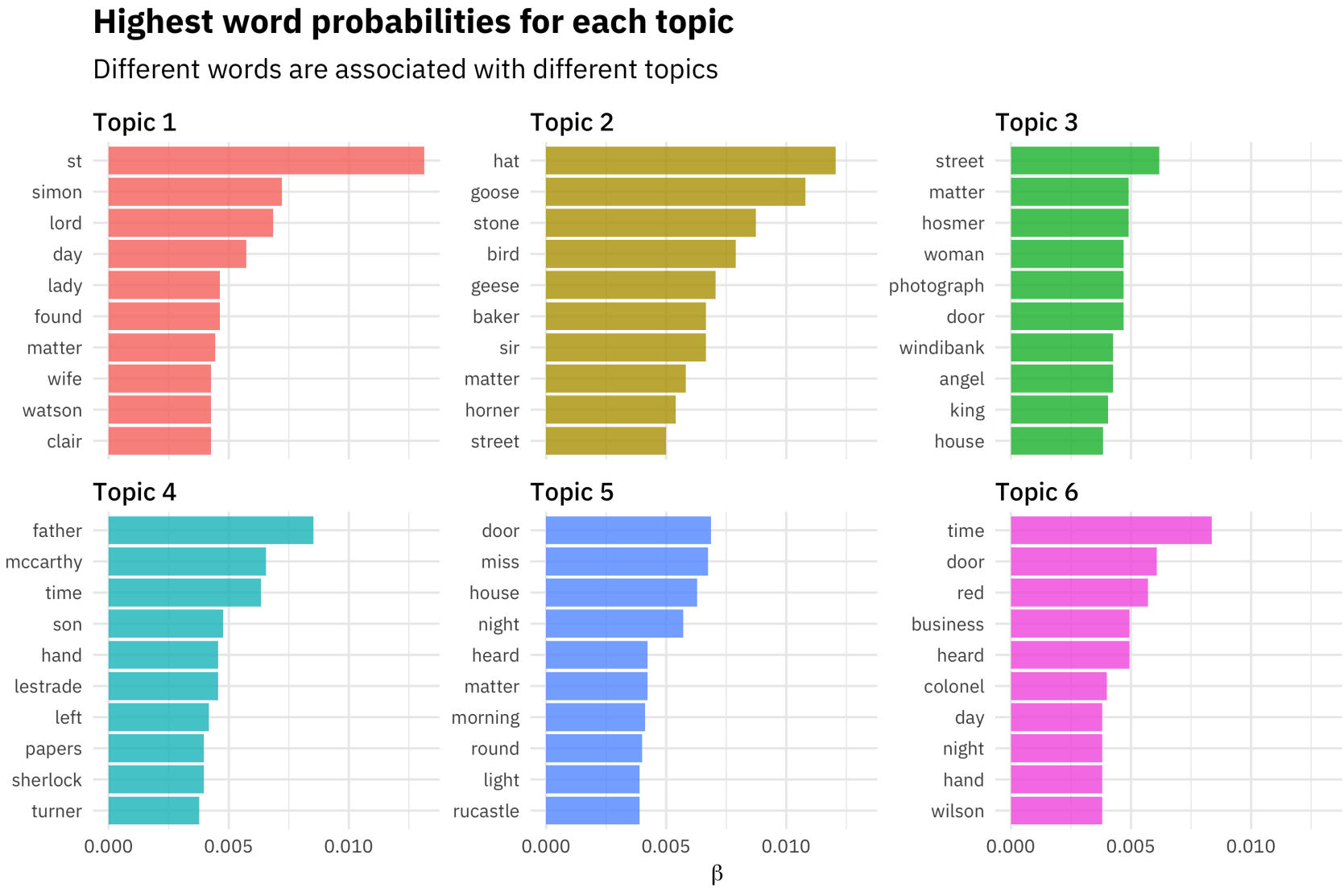
Performant unsupervised language models

$\Downarrow$

cheap applications leveraging language **understanding**.

```
This is one of Crichton's best books. The characters of Karen Ross, Peter Elliot,
Munro, and Amy are beautifully developed and their interactions are exciting,
complex, and fast-paced throughout this impressive novel. And about 99.8 percent of
that got lost in the film. Seriously, the screenplay AND the directing were
horrendous and clearly done by people who could not fathom what was good about the
novel. I can't fault the actors because frankly, they never had a chance to make this
turkey live up to Crichton's original work. I know good novels, especially those with
a science fiction edge, are hard to bring to the screen in a way that lives up to the
original. But this may be the absolute worst disparity in quality between novel and
screen adaptation ever. The book is really, really good. The movie is just dreadful.
```

Language understanding at scale: Topic models



The history of Distributional Semantics

You shall know a word by the company it keeps!

(Firth, 1957)

You shall know a word by the company it keeps!

(Firth, 1957)

- "In most cases, the meaning of a word is its use in language." (Wittgenstein and Anscombe, 1954)

You shall know a word by the company it keeps!

(Firth, 1957)

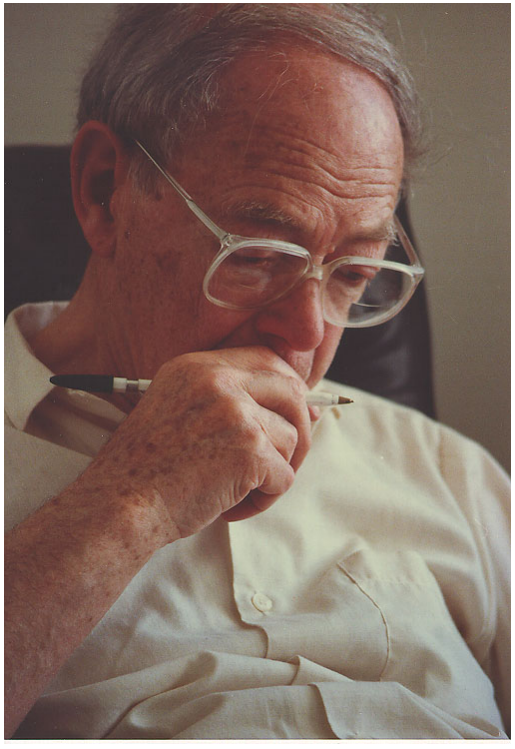
- "In most cases, the meaning of a word is its use in language." (Wittgenstein and Anscombe, 1954)
- "Words which are similar in meaning occur in similar contexts" (Rubenstein and Goodenough, 1965)

You shall know a word by the company it keeps!

(Firth, 1957)

- "In most cases, the meaning of a word is its use in language." (Wittgenstein and Anscombe, 1954)
- "Words which are similar in meaning occur in similar contexts" (Rubenstein and Goodenough, 1965)
- "Two words are semantically similar to the extent that their contextual representations are similar" (Miller and Charles, 1991)

Distributional Structure, (Harris, 1954)



Distributional Structure, (Harris, 1954)

Distributional Hypothesis Language can be understood in terms of the occurrence of parts relative to other parts.

The distribution of an element is the sum of all its environments.

The distributional hypothesis reified

Distributional Semantic Models.

Count based models have existed since the 70s, largely originating from information retrieval.

Idea Model documents by the terms that occur in them.

Term-document matrix

	D1	D2	D3
good	2	0	10
bad	7	0	0
awesome	0	4	0
blue	1	0	4
green	0	1	0

Term-document matrix

	D1	D2	D3
good	2	0	10
bad	7	0	0
awesome	0	4	0
blue	1	0	4
green	0	1	0

What about synonymous or similar words?

Factorized Term-document matrix

	L1	L2
D1	0.3	0.6
D2	0.2	0.1
D3	0.1	0.5

Factorized term-document matrix

	L1	L2
good	0.1	0.6
bad	0.3	0.7
awesome	0.2	0.5
blue	0.8	0.2
green	0.7	0.2

Latent Dirichlet Allocation

Assumes a generative process:

- Choose document length $N \sim$ Poisson
- Choose topic distribution $\Theta \sim$ Dirichlet
- For each word,
  - Choose a topic $z \sim$ Multinomial(Θ)
  - Choose a word from the topic.

Term-term models

Idea Model terms directly by the terms they occur with.

Term-term matrix

	bank	interest	finals
cash	300	210	133
sport	75	140	200

Semantic Neighborhoods

Word	Neighbors
absolutely	absurd, whatsoever, totally, exactly, nothing, does . . .
bottomed	dip, copper, drops, topped, slide, trimmed, slightly . . .
captivating	shimmer, stunningly, superbly, plucky, witty . . .
doghouse	dog, porch, crawling, beside, downstairs, gazed . . .

What do we capture?

Paradigmatic relations Interchangeability relations, e.g. *dog* - *cat*
Syntagmatic relations Joint usage relations, e.g. *dog* - *bone*

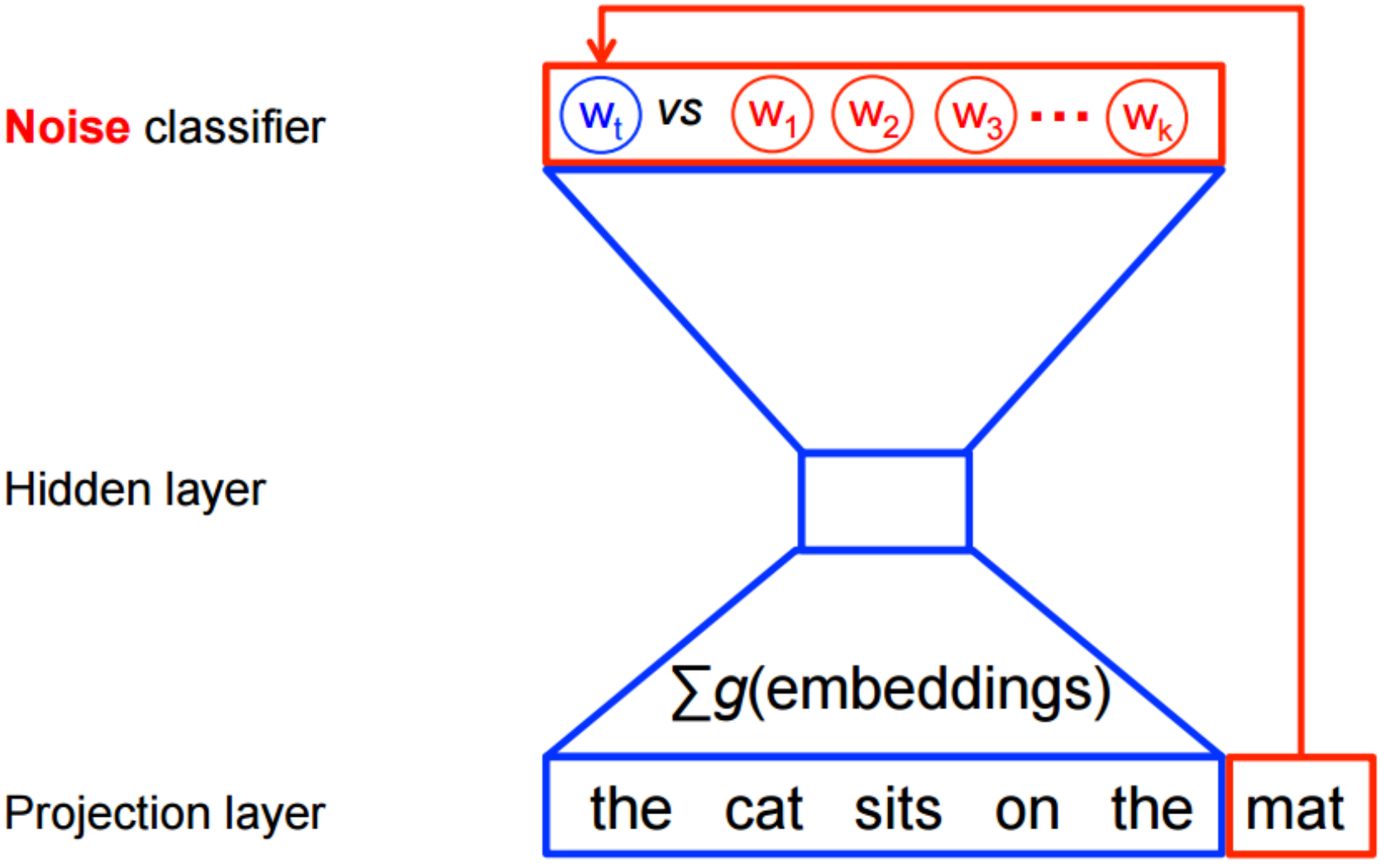
What do we capture?

Paradigmatic relations Interchangeability relations, e.g. *dog* - *cat*
Syntagmatic relations Joint usage relations, e.g. *dog* - *bone*

Depends on the model specification **and** our mode of query.

Don't count, Predict!

Don't count, Predict!



Current Language Models

Shallow models

- Word2Vec

Shallow models

- Word2Vec
  - Fasttext

Shallow models

- Word2Vec
  - Fasttext
  - Doc2Vec

Shallow models

- Word2Vec
  - Fasttext
  - Doc2Vec
- GloVe

Word2Vec (CBoW)

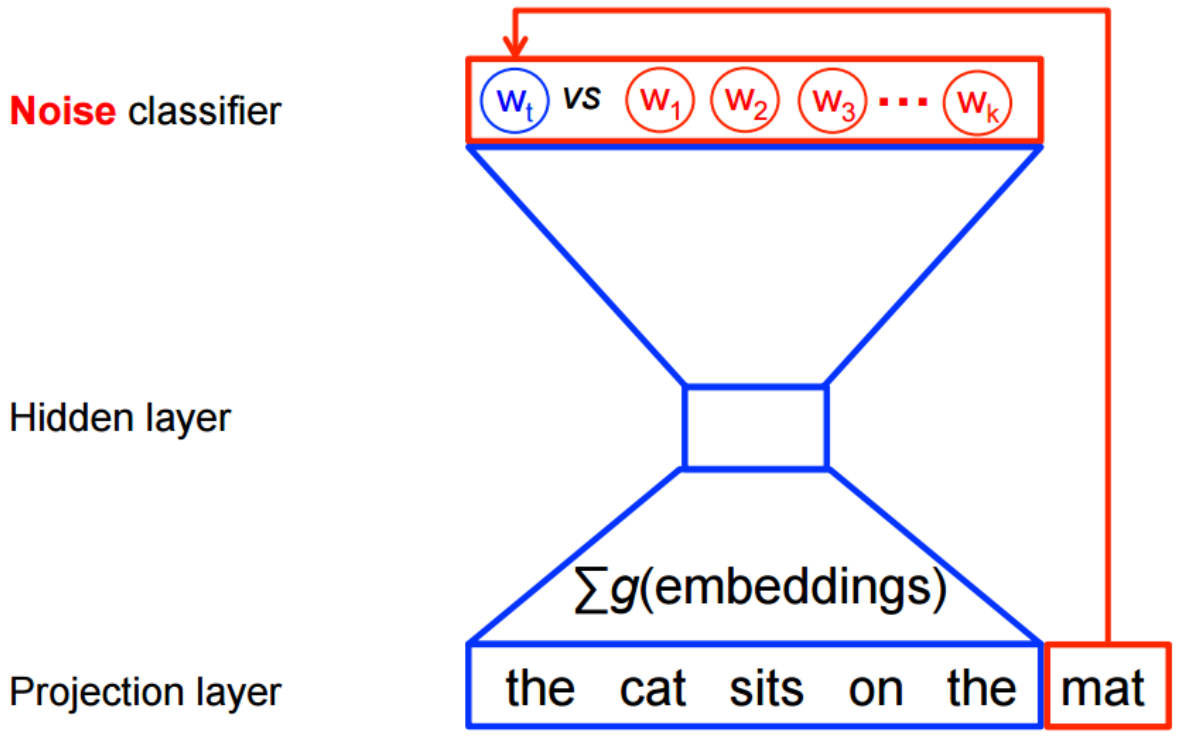
Learn to predict a word in context (Cloze test)

```
Today, I went to the _____ and bought some milk and egg
```

Word2Vec (CBoW)

Learn to predict a word in context (Cloze test)

```
Today, I went to the _____ and bought some milk and egg
```



Word2Vec

Word2Vec

$V \sim$ Vocabulary

$T : V \to \mathbb{R}^D$

$C : V \to \mathbb{R}^D$

$w_t = \langle T(\text{mat}), C(\text{the}) + C(\text{cat}) + \ldots \rangle$

Word2Vec

$V \sim$ Vocabulary

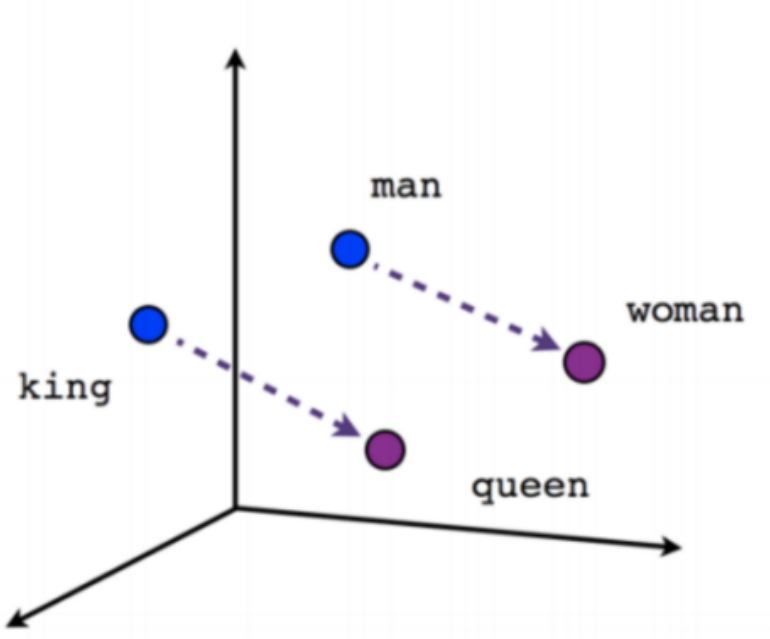
$T : V \to \mathbb{R}^D$

$C : V \to \mathbb{R}^D$

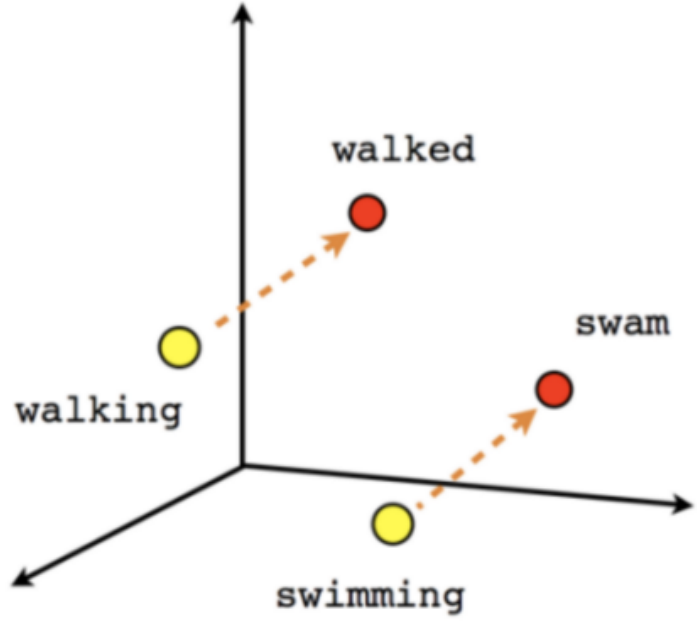
$w_t = \langle T(\text{mat}), C(\text{the}) + C(\text{cat}) + \ldots \rangle$

$\implies$ dot product between $C(x)$ and $T(y)$ relates to the probability of observing x in the context of y.

Word2Vec



Male-Female



Verb tense

Term-? models

Prediction based models are much easier to extend than count based models.

- subword representations (FastText)

Term-? models

Prediction based models are much easier to extend than count based models.

- subword representations (FastText)
- positional masking (FastText)

Term-? models

Prediction based models are much easier to extend than count based models.

- subword representations (FastText)
- positional masking (FastText)
- document representations (Doc2Vec)

Quiz-time

Given the following data:

```
A    X    P
A    Y    Q

B    X    P
B    Y    Q

C    X    Q
C    Y    P
```

- Which item is more similar to A, B or C?

Quiz-time

Given the following data:

```
A    X    P
A    Y    Q

B    X    P
B    Y    Q

C    X    Q
C    Y    P
```

- Which item is more similar to A, B or C?
- Can count based term-term models discriminate between A, B, and C?

Quiz-time

Given the following data:

```
A    X    P
A    Y    Q

B    X    P
B    Y    Q

C    X    Q
C    Y    P
```

- Which item is more similar to A, B or C?
- Can count based term-term models discriminate between A, B, and C?
- Can Word2Vec models discriminate between A, B, and C?

Deep models

- Recurrent Neural Networks

Deep models

- Recurrent Neural Networks
- ELMo

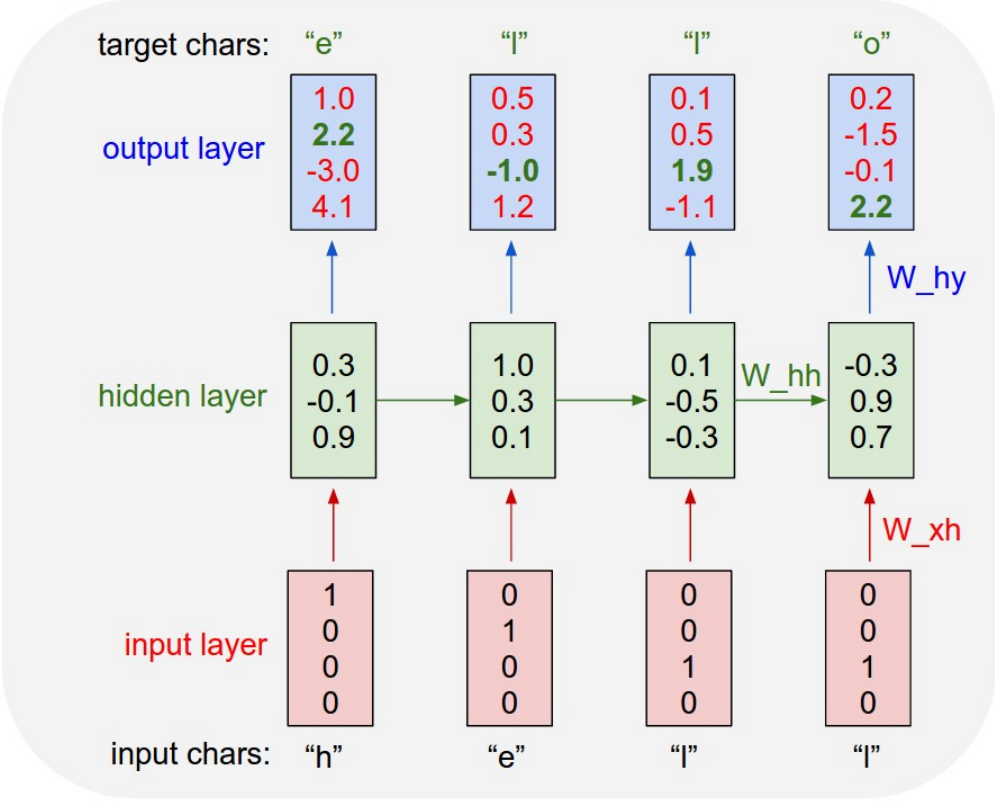
Deep models

- Recurrent Neural Networks
- ELMo
- BERT

Recurrent Neural Networks

Next step prediction:

```
Today, I went to the _____
Today, I went to the store ___
Today, I went to the store and ____
```



ELMo

- 2018
- Next step prediction.
- Attention
- Two recurrent networks, one forward-predicting and one backwards-predicting.
- Contextualized word embedding: The context influences the representation of terms.

BERT

- Cloze test. (masks 15 % of the input terms)
- Attention
- Transformer network. (Deep Feed forward)
- Contextualized word embeddings: The context, again, influences the representation of terms.

Transfer learning

Using representations from contextualized, self-supervised, language models we can solve many supervised problems much quicker.

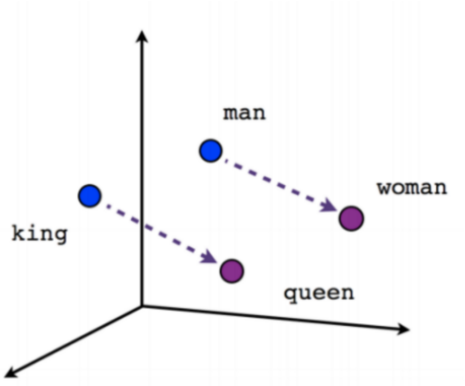
System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Future

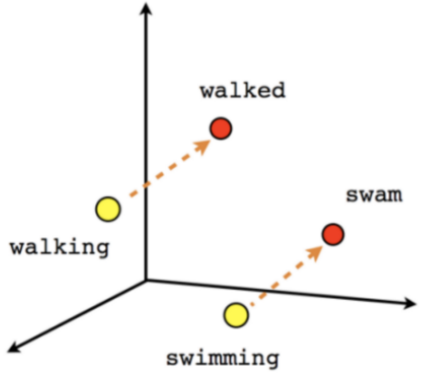
- New, even better, language models
- Wider application areas
- **All** models I've talked about are unsupervised (or self-supervised).

Using Language Models

What do they contain?

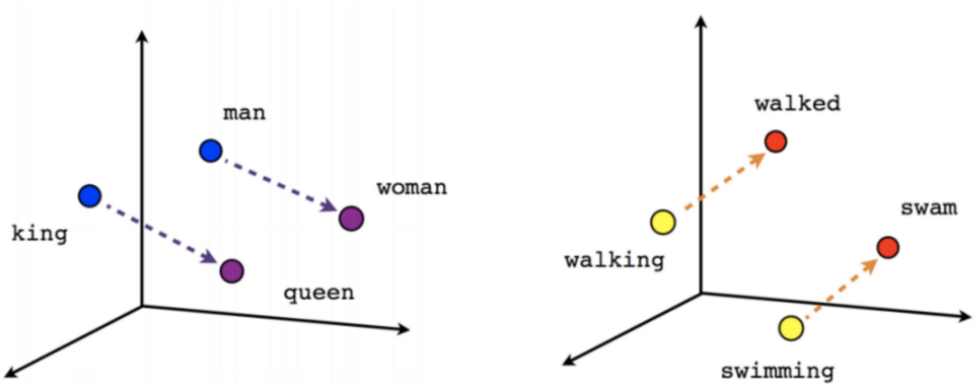


Male-Female



Verb tense

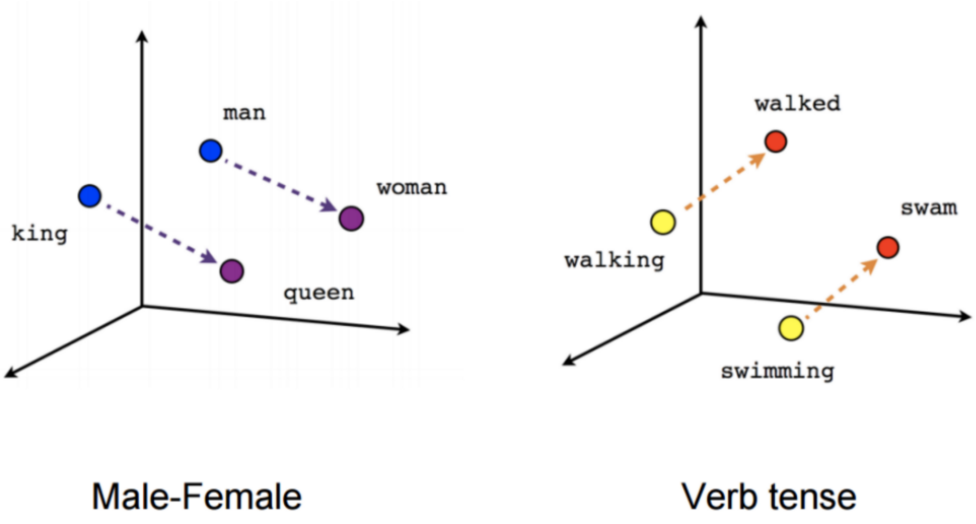
What do they contain?



Male-Female Verb tense

Bias in word embeddings "man is to computer programming as woman is to homemaker"

What do they contain?



Male-Female Verb tense

Bias in word embeddings "man is to computer programming as woman is to homemaker"

Misogyny, racism, et.c. will be as present in the models as they are in the data.

Problem or Opportunity?

- We could argue that bias in embeddings corresponds to real world bias. i.e. automated bias detection?

Problem or Opportunity?

- We could argue that bias in embeddings corresponds to real world bias. i.e. automated bias detection?
- How are people talking about democracy? Or corruption?

Language models in networks

- How does language use differ between various sources? When talking about “transformers” do AI-researches and electrician talk about the same thing?
- What is being communicated in which points in the network?
- graph2vec

Responsibilities

GPT-2

"This is the story of a little dog named Boo." "This is the story of a little dog named Boo. You see, Boo is a Labrador, a dog that loves people more than anything else. A little over a year ago, Boo became a part of The Dog Rescue of Arlington, Virginia. Boo was a puppy (now 2 years old) that had been abandoned. Boo's story was heartbreaking."

Responsibilities

GPT-2

"This is the story of a little dog named Boo." "This is the story of a little dog named Boo. You see, Boo is a Labrador, a dog that loves people more than anything else. A little over a year ago, Boo became a part of The Dog Rescue of Arlington, Virginia. Boo was a puppy (now 2 years old) that had been abandoned. Boo's story was heartbreaking."

Decided **not** to release their complete model, for fear of malicious use.