Natural Language Processing

Johan Boye, KTH

September 18, 2020

Natural Language Processing (NLP)

What is a “natural” language?

- English, Swedish, Chinese, Russian, Arabic, ...

What is “processing”?

- One-liner: “Making computers understand language”
- More precisely: Extracting useful information from natural language, and generating (correct, useful) language.

What’s the point?

Many useful applications:

- Speech recognition
- Machine translation
- Search engines
- Natural-language interfaces
- Question-answering
- Text summarization
- Text classification
- Word prediction
- Spell checking
- Grammar checking
- ...

NLP and Artificial Intelligence

Remember the Turing Test?

- Language understanding as an AI litmus test

Language is a *distinctive* and *essential* part of human intelligence

Language understanding is intimately connected to other AI topics:

- Planning, reasoning, knowledge representation, ...

Why is NLP hard?

Language is *underspecified*

Language interpretation *requires knowledge about the world*.

Language is *ambiguous*

- “John made the pie in the fridge.”

+ many more (smaller) challenges...

Computational models of language

4 computational models:

- Language as a *bag of words*
- Language as a *probabilistic sequence*
- Language as a *tree structure*
- Language as a *vector space*

Language as a bag of words

Observation: Language consists of words

Can we use (only) this observation for doing useful applications?

To Sherlock Holmes she is always *the* woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex. It was not that he felt any emotion akin to love for Irene Adler. All emotions, and that one particularly, were abhorrent to his cold, precise but admirably balanced mind. He was, I take it, the most perfect reasoning and observing machine that the world has seen, but as a lover he would have placed himself in a false position. He never spoke of the softer passions, save with a gibe and a sneer. They were admirable things for the observer—excellent for drawing the veil from men's motives and actions. But for the trained reasoner to admit such intrusions into his own delicate and finely adjusted temperament was to introduce a distracting factor which might throw a doubt upon all his mental results. Grit in a sensitive instrument, or a crack in one of his own high-power lenses, would not be more disturbing than a strong emotion in a nature such as his. And yet there was but one woman to him, and that woman was the late Irene Adler, of dubious and questionable memory.

To Sherlock Holmes she is always *the* woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex. It was not that he felt any emotion akin to love for Irene Adler. All emotions, and that one particularly, were abhorrent to his cold, precise but admirably balanced mind. He was, I take it, the most perfect reasoning and observing machine that the world has seen, but as a lover he would have placed himself in a false position. He never spoke of the softer passions, save with a gibe and a sneer. They were admirable things for the observer—excellent for drawing the veil from men's motives and actions. But for the trained reasoner to admit such intrusions into his own delicate and finely adjusted temperament was to introduce a distracting factor which might throw a doubt upon all his mental results. Grit in a sensitive instrument, or a crack in one of his own high-power lenses, would not be more disturbing than a strong emotion in a nature such as his. And yet there was but one woman to him, and that woman was the late Irene Adler, of dubious and questionable memory.

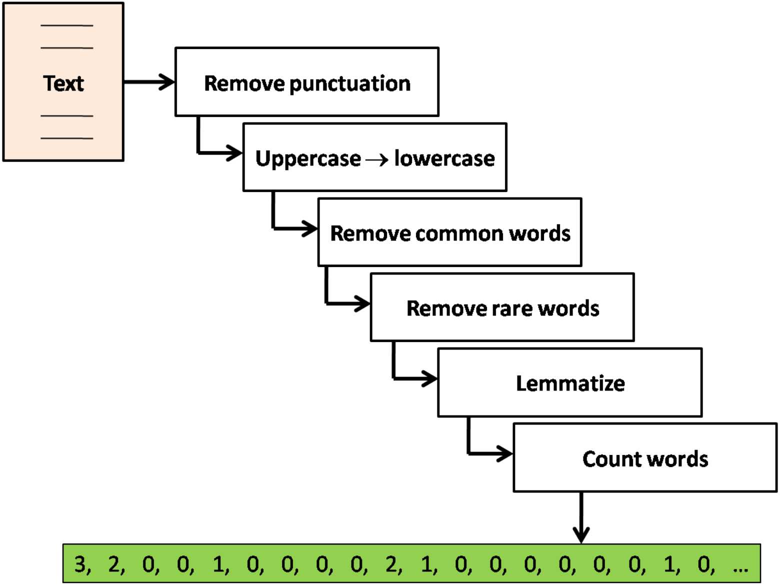
10	a
9	the
9	and
6	to
6	his
5	was
5	in
4	that
4	of
4	he
4	for
4	but
3	woman
3	one
2	would
2	were
2	such
2	she
2	own
2	not

To **Sherlock Holmes** she is **always** *the* **woman**. I have **seldom heard** him **mention** her under any other **name**. In his **eyes** she **eclipses** and **predominates** the **whole** of her **sex**. It was not that he **felt** any **emotion akin** to **love** for **Irene Adler**. All **emotions**, and that one **particularly**, were **abhorrent** to his **cold**, **precise** but **admirably balanced mind**. He was, I **take** it, the **most perfect reasoning** and **observing machine** that the **world** has **seen**, but as a **lover** he would have **placed himself** in a **false position**. He **never spoke** of the **softer passions**, **save** with a **gibe** and a **sneer**. They were **admirable things** for the **observer**—**excellent** for **drawing** the **veil** from **men's motives** and **actions**. But for the **trained reasoner** to **admit such intrusions** into his **own delicate** and **finely adjusted temperament** was to **introduce** a **distracting factor** which might **throw** a **doubt** upon all his **mental results**. **Grit** in a **sensitive instrument**, or a **crack** in one of his **own high-power lenses**, would not be **more disturbing** than a **strong emotion** in a **nature** such as his. And **yet** there was but one **woman** to him, and that **woman** was the **late Irene Adler**, of **dubious** and **questionable memory**.

3 woman
2 own
2 irene
2 emotion
2 adler
1 yet
1 world
1 whole
1 veil
1 trained
1 throw
1 things
1 the
1 temperament
1 take
1 such
1 strong
1 spoke
1 softer
1 sneer

To **Sherlock Holmes** she is **always** *the* **woman**. I have **seldom hear** him **mention** her under any other **name.** In his **eye** she **eclipse** and **predominate** the **whole** of her **sex**. It was not that he **feel** any **emotion akin** to **love** for **Irene Adler**. All **emotion** , and that one **particularly**, were **abhorrent** to his **cold**, **precise** but **admirably balanced mind.** He was, I **take** it, the **most perfect reasoning** and **observing machine** that the **world** has **see** , but as a **lover** he would have **place himself** in a **false position.** He **never speak** of the **soft passion** , **save** with a **gibe** and a **sneer**. They were **admirable things** for the **observer**— **excellent** for **draw** the **veil** from **man motive** and **action** . But for the **train reasoner** to **admit such intrusion** into his **own delicate** and **finely adjust temperament** was to **introduce** a **distracting factor** which might **throw** a **doubt** upon all his **mental result** . **Grit** in a **sensitive instrument**, or a **crack** in one of his **own high-power lens** , would not be **more disturbing** than a **strong emotion** in a **nature** such as his. And **yet** there was but one **woman** to him, and that **woman** was the **late Irene Adler**, of **dubious** and **questionable memory**.

3	woman
2	own
2	irene
2	emotion
2	adler
1	yet
1	world
1	whole
1	veil
1	train
1	throw
1	things
1	the
1	temperament
1	take
1	such
1	strong
1	speak
1	soft
1	sneer



Text

Remove punctuation

Uppercase → lowercase

Remove common words

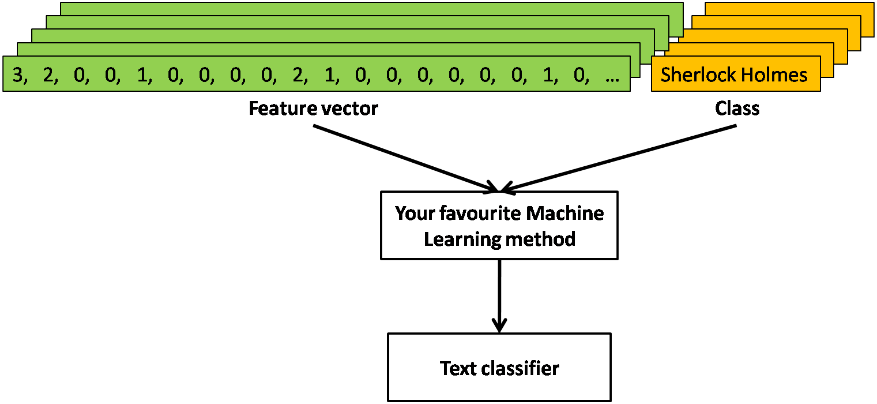
Remove rare words

Lemmatize

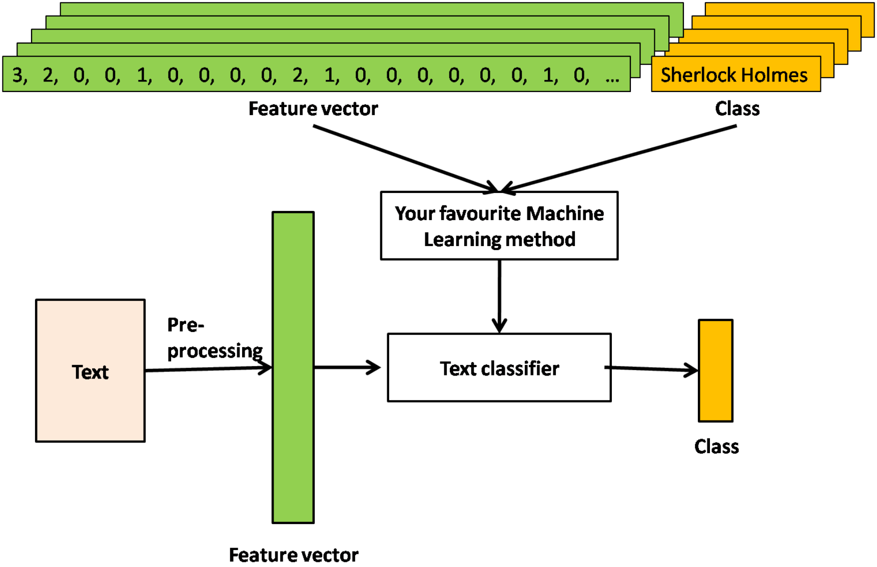
Count words

3, 2, 0, 0, 1, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 1, 0, ...

Feature vector



3, 2, 0, 0, 1, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 1, 0, ...



3, 2, 0, 0, 1, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 1, 0, ...

Sherlock Holmes

Feature vector

Class

Your favourite Machine Learning method

Text

Pre-processing

Text classifier

Class

Feature vector

Language as a bag of words

The bag-of-words model is useful for *text classification*:

- Spam detection
- Sentiment analysis
- Author identification
- ...

Language as a probabilistic sequence

Observations:

- Language consists of *sequences* of words.
- Not all sequences are equally probable.

- *I would like to*
- ... is probable.
- *Furiously sleep ideas green colorless*
- ... is improbable

Counting bigrams

. . . $\boxed{\text{I would}}$ like to know your plans for . . .

I would: 1

Counting bigrams

. . . I $\boxed{\text{would like}}$ to know your plans for . . .

I would: 1   would like: 1

Counting bigrams

. . . I would $\boxed{\text{like to}}$ know your plans for . . .

I would: 1 would like: 1 like to: 1

Bigram probabilities

The word *I* occurred 1000 times...

- ... 20 times, the next word was *like*
- ... 200 times, the next word was *am*
- ... 100 times, the next word was *have*
- etc.

From the counts, we can estimate *bigram probabilities*:

- $P(like|I) = 0.02$
- $P(am|I) = 0.2$
- $P(have|I) = 0.1$
- etc.

Trigram probabilities

The sequence *I like* occurred 20 times...

- ... 5 times, the next word was *to*
- ... 4 times, the next word was *that*
- ... 1 time, the next word was *apples*
- etc.

From the counts, we can estimate *trigram probabilities*:

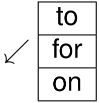
- $P(\textit{to}|\textit{I like}) = 0.25$
- $P(\textit{that}|\textit{I like}) = 0.2$
- $P(\textit{apples}|\textit{I like}) = 0.05$
- etc.

n-gram probabilities

- In general, *n*-gram probabilities: $P(w_n|w_1, \ldots, w_{n-1})$
- Given that we've seen the sequence $w_1, \ldots, w_{n-1}$, what is the probability that we will now see w_n?
- Higher $n \Rightarrow$ we capture more language structure, BUT
- Higher $n \Rightarrow$ we need more training data to get accurate probabilities.
- 4-grams and above require Google quantities of data, OR a restricted domain!

Applications of *n*-gram models

- Word prediction
  - *I'm going* ______



to

for

on

- Spelling correction
  - *Flights form Boston.*
- Speech recognition
  - $P(\text{“recognize speech”}) > P(\text{“wreck a nice beach”})$
- Translation
  - $P(\text{“tall building”}) > P(\text{“high building”})$
- ... and many more

Limitations of *n*-gram models

These two sentences are highly improbable:

- *Colorless green ideas sleep furiously*
- *Furiously sleep ideas green colorless* *

Yet the first one is correct, and the second is incorrect!

Why is this, and how do we instinctively know it?

Limitations of *n*-gram models

Some correct sentences:

- *Either* Tom sleeps, *or* he eats.
- *If* Tom sleeps, *then* he dreams.
- *If either* Tom sleeps *or* he eats, *then either* Tom dreams *or* he is happy.

Some incorrect sentences:

- *Either* Tom sleeps, *then* he eats. *
- *If* Tom sleeps, *or* he dreams. *
- *If either* Tom sleeps *or* he eats, *or either* Tom dreams *or* he is happy. *

Limitations of *n*-gram models

n-gram models can NOT:

- ... separate correct and incorrect sentences.
- ... capture long-distance dependencies.
- ... represent relationships between words.

Noam Chomsky pointed all this out in a ground-breaking book (*Syntactic structures*) in 1956. He also presented a solution...

Language as a tree structure

Chomsky invented the *context-free grammar*:

$\boxed{S \rightarrow NP\ VP}$

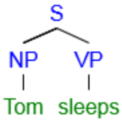
"A sentence is a noun phrase followed by a verb phrase."

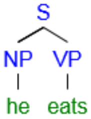
$\boxed{NP \rightarrow Name \mid PN}$

"A noun phrase is a name, or a pronoun."

$\boxed{VP \rightarrow Verb \mid Verb\ Noun}$

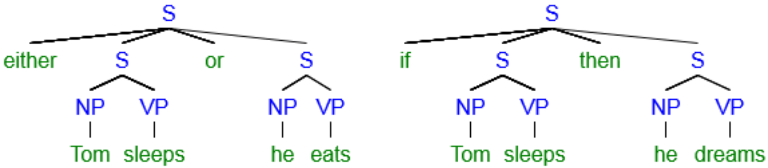
"A verb phrase is a verb, or a verb followed by a noun."





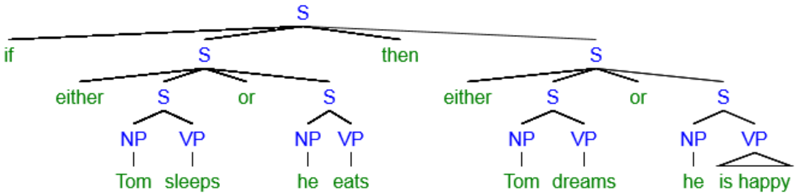
Language as a tree structure

$$
\begin{aligned}
S &\rightarrow \text{ either } S \text{ or } S \\
S &\rightarrow \text{ if } S \text{ then } S
\end{aligned}
$$



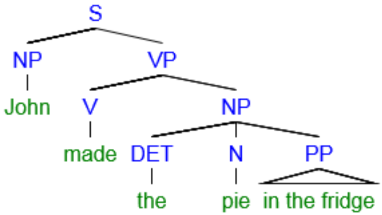
Language as a tree structure

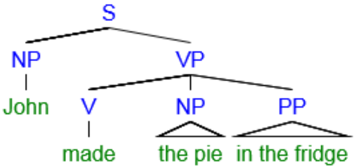
$$\begin{aligned} S &\rightarrow \text{ either } S \text{ or } S \\ S &\rightarrow \text{ if } S \text{ then } S \end{aligned}$$



Tree structure and ambiguities

“John made the pie in the fridge”





Intended interpretation ⇑

Language as a tree structure

Tree-based approaches...

- ... can separate correct and incorrect sentences.
- ... can tell you relations between phrases, even if they are not adjacent.
- ... can represent different interpretations of the same sentence.

Useful for Information extraction, Question-answering, etc.

Language and meaning

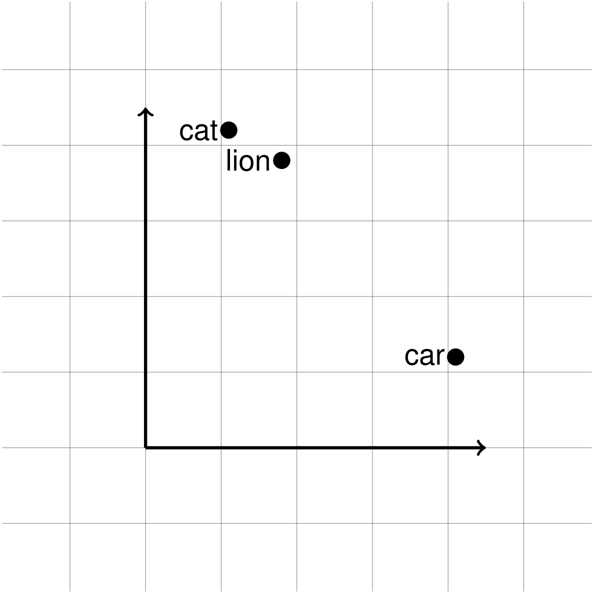
Observation: Words have *meaning*.

What is the meaning of *cat*?

Language and meaning

- What is the meaning of *cat*?
- Surprisingly hard to answer, but...
- ... the meaning of *cat* should be more similar to the meaning of *lion*...
- ... than to the meaning of *car*.
- Idea: View words as points or vectors in a high-dimensional vector space.

Language as a vector space



Word-document matrix

One way of creating word vectors is through a *word-document matrix*.

		Documents						
		1	2	3	4	5	. . .	*n*
	aalborg	1						
	aback		1		1			
	abandon		1					
Words	. . .							
	zombie					1		
	zone							
	zurich			1				

Word-document matrix

The word vector for *aback* is highlighted below.

		Documents						
		1	2	3	4	5	. . .	*n*
	aalborg	1						
	aback	**0**	**1**	**0**	**1**	**0**	. . .	**0**
	abandon		1					
Words	. . .			. . .				
	zombie					1		
	zone							
	zurich			1				

Word2Vec

Imagine (again) a context window sliding over text:

$\ldots$ I would like $\underbrace{\textbf{to}}_{\text{focus word}}$ know the plans for $\ldots$

Then: Train a neural network that can predict the focus word from the context words.

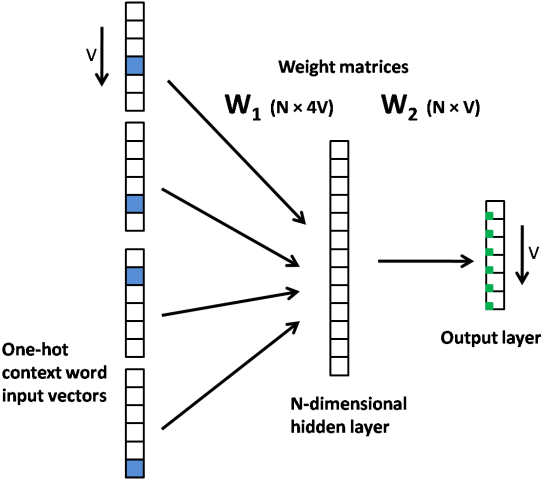
Word2Vec training

For input and output to the network, we use a *1-hot-encoding*, e.g.

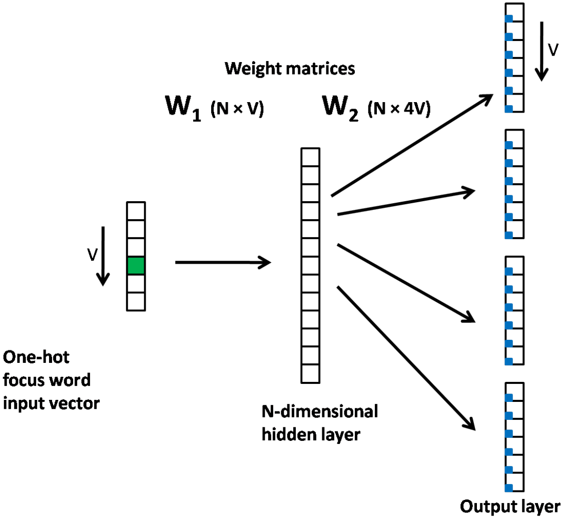
would	(**1**00000. . .0)
like	(0**1**0000. . .0)
to	(00**1**000. . .0)
know	(000**1**00. . .0)
the	(0000**1**0. . .0)

Each vector has length V (= size of the vocabulary).

Continuous-bag-of-words model


V
Weight matrices
W_1 (N × 4V)
W_2 (N × V)
V
Output layer
One-hot context word input vectors
N-dimensional hidden layer

Skipgram model



Word2Vec skipgram model

After training on *lots* of text, column i in the weight matrix W_1 contains a word vector for the ith word.

$$W = \begin{pmatrix} 0.243 & \text{-0.756} & 0.169 & -0.564 & 0.012 \\ -0.912 & \text{-0.213} & 0.442 & -0.767 & 0.851 \\ & & \dots & & \\ 0.516 & 0.760 & -0.109 & 0.926 & -0.429 \end{pmatrix}$$

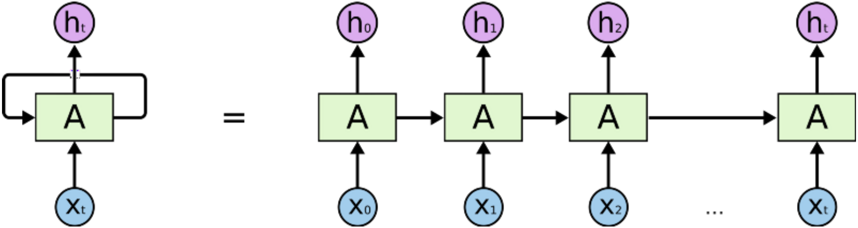
🟦 = word vector for word 2 (e.g. *like*).

Word vector relations

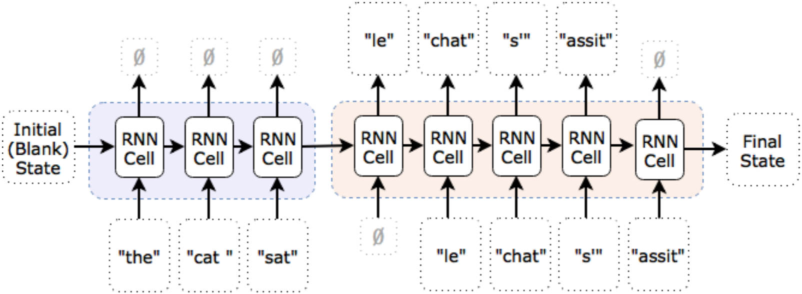
Word vectors trained with Word2Vec have many amazing features:

$$v(\mathit{France}) - v(\mathit{Paris}) + v(\mathit{Rome}) = v(\mathit{Italy})$$
$$v(\mathit{king}) - v(\mathit{man}) + v(\mathit{woman}) = v(\mathit{queen})$$

Recurrent neural networks (RNNs)



Encoder-decoder architecture

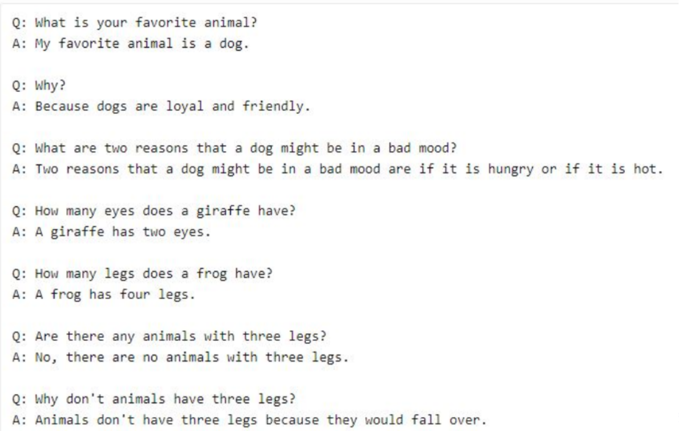


GPT-3: Is it ’real’ AI?

GPT-3 is a language model released by OpenAI in May 2020.

Trained on a huge dataset, GPT-3 has 175 **billion** parameters.

Already many amazing demos have been built on top of GPT-3, like this QA system:



```
Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many legs does a frog have?
A: A frog has four legs.

Q: Are there any animals with three legs?
A: No, there are no animals with three legs.

Q: Why don't animals have three legs?
A: Animals don't have three legs because they would fall over.
```

Sum-up

Natural Language Processing is:

- a “hot” research topic, with many useful applications
- very relevant for Artificial Intelligence
- a good career choice if you like languages AND computer programming AND maths

Learn more...

- DD2418 Language Engineering (period 4)
- DD2476 Search Engines and Information Retrieval Systems (period 3 and 4)