Recap

- Word2Vec
  - Method to represent words in vector space
  - Requires large quantity of raw text
  - Pre-trained embeddings can be shared
  - Embeddings capture (some aspects of) lexical meaning

Large Language Models

Sprachverarbeitung (VL + Ü)

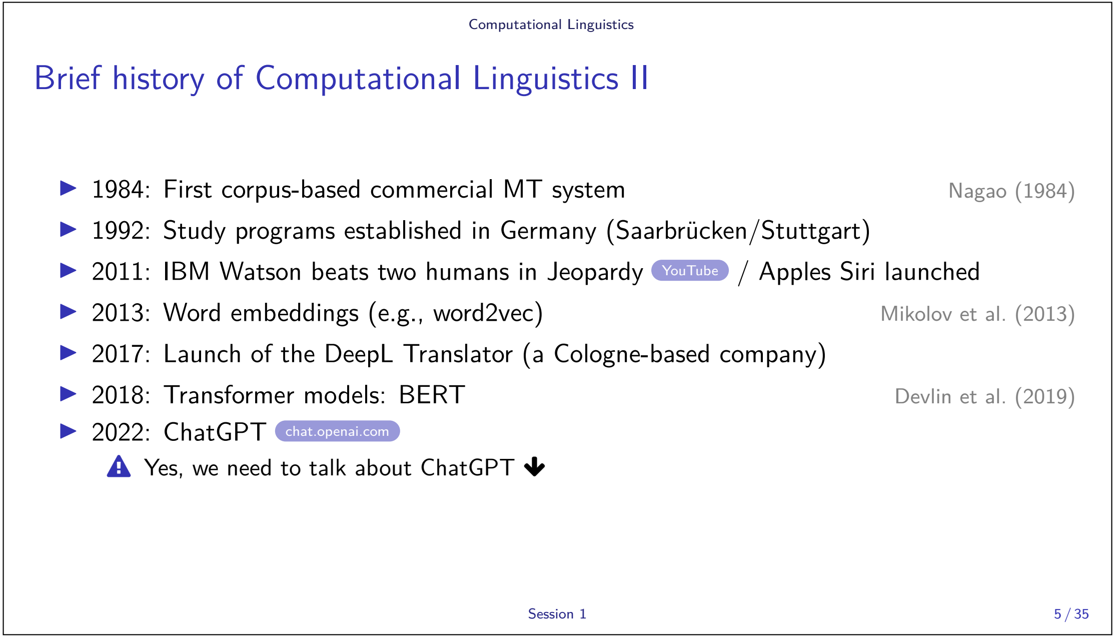
Nils Reiter

June 29, 2023

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Group Exercise

1. In which situations have you talked about ChatGPT (& co)?
2. For which tasks can it be put to use?
3. For which tasks *should* it not be used? Why not?



Brief history of Computational Linguistics II

- 1984: First corpus-based commercial MT system — Nagao (1984)
- 1992: Study programs established in Germany (Saarbrücken/Stuttgart)
- 2011: IBM Watson beats two humans in Jeopardy YouTube / Apples Siri launched
- 2013: Word embeddings (e.g., word2vec) — Mikolov et al. (2013)
- 2017: Launch of the DeepL Translator (a Cologne-based company)
- 2018: Transformer models: BERT — Devlin et al. (2019)
- 2022: ChatGPT chat.openai.com
  - ⚠ Yes, we need to talk about ChatGPT ↓

Large Language Models

- Term LLM used in contrast to classical language models
- Family of »transformer models«: BERT, GPT, ...
  - BERT by Google, GPT-X by OpenAI
  - BERT model can be downloaded and used locally
- Huge amount of training data (e.g., the web)
- High computing costs
  - »Just how much does it cost to train a model? Two correct answers are ›depends‹ and ›a lot‹.« Sharir et al. (2020, 1)
  - BERT w/ 340 million parameters: $ 10k / $ 200k

Key Idea 1: Learned Representation

- Classical ML: Instances are represented by their features
- Neural ML
  - Words/texts are represented by vectors
  - Vectors are learned representations
    - I.e., vectors are optimised for some task, usually filling gaps in texts

Key Idea 1: Learned Representation

- Classical ML: Instances are represented by their features
- Neural ML
  - Words/texts are represented by vectors
  - Vectors are learned representations
    - I.e., vectors are optimised for some task, usually filling gaps in texts

$$\boxed{p\left(w_n \,\middle|\, \begin{bmatrix} w_1 \\ \text{wortart}_1 \\ \vdots \end{bmatrix}, \begin{bmatrix} w_2 \\ \text{wortart}_2 \\ \vdots \end{bmatrix}, \cdots\right)} \longrightarrow \boxed{p\left(w_n \,\middle|\, \begin{bmatrix} 0.105 \\ -0.145 \\ 0.742 \\ \vdots \end{bmatrix}, \begin{bmatrix} -0.427 \\ 0.864 \\ -0.232 \\ \vdots \end{bmatrix}, \cdots\right)}$$

Linguistic context properties

Properties have been learned from data

Key Idea 2: Not every token is equally important

- »Attention Is All You Need« Vaswani et al. (2017)
- Idea: During training, model learns which tokens are relevant to predict the output
  - Additional parameters to train ...

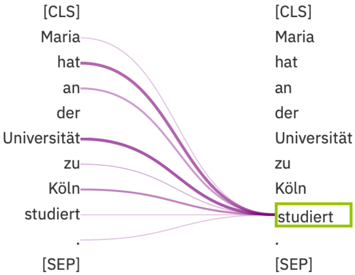


Figure: Attention given when predicting the word »studiert«. Screenshot taken from `https://huggingface.co/spaces/exbert-project/exbert`

Key Idea 3: Training Process Split into two Phases

- Traditionally (Naive Bayes, Decision tree, ...), we train a model and are done
- Transformer architecture:
  - Pre-Training: Model is trained on huge data set to do generic task
  - Fine-Tuning: We continue training the model, but on the task we are actually interested in (!)

Key Idea 3: Training Process Split into two Phases

BERT Training Tasks

Masked Language Modeling (MLM)

- Sentence-wise
- 15% of the tokens are »masked« by a special token
- Model predicts these, having access to all other tokens

Key Idea 3: Training Process Split into two Phases

BERT Training Tasks

Masked Language Modeling (MLM)

- Sentence-wise
- 15% of the tokens are »masked« by a special token
- Model predicts these, having access to all other tokens

Next sentence prediction (NSP)

- Two (masked) sentences are concatenated
- Model has to predict wether second sentence follows on the first or not

Key Idea 4: Scale Up

- With the transformer recipe, many parameters have simply been scaled up

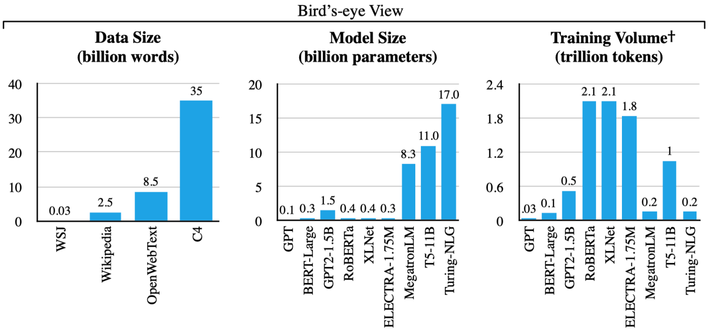
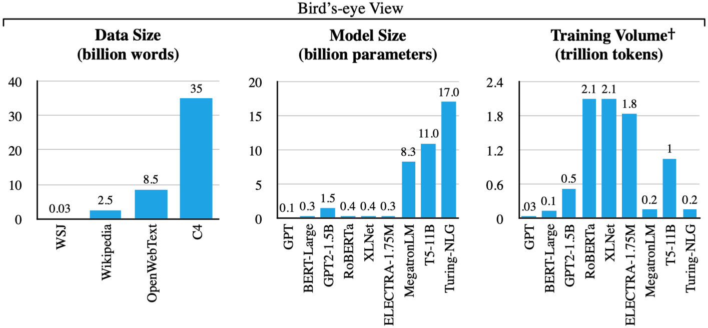
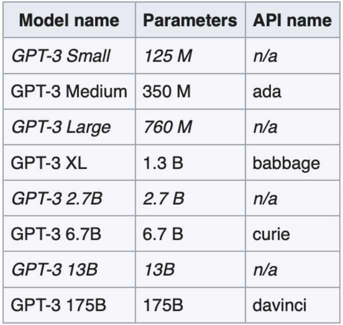


Figure: Statistics about NLP models (Sharir et al., 2020; Wikipedia)

Key Idea 4: Scale Up

- With the transformer recipe, many parameters have simply been scaled up



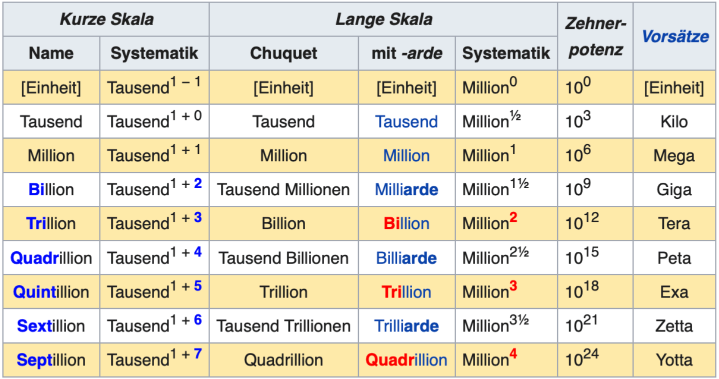


Model name	Parameters	API name
GPT-3 Small	*125 M*	*n/a*
GPT-3 Medium	350 M	ada
GPT-3 Large	*760 M*	*n/a*
GPT-3 XL	1.3 B	babbage
GPT-3 2.7B	*2.7 B*	*n/a*
GPT-3 6.7B	6.7 B	curie
GPT-3 13B	*13B*	*n/a*
GPT-3 175B	175B	davinci

Figure: Statistics about NLP models (Sharir et al., 2020; Wikipedia)

Key Idea 4: Scale Up

Large Numbers are Complicated



Kurze Skala		*Lange Skala*			*Zehner-potenz*	*Vorsätze*
Name	**Systematik**	**Chuquet**	**mit *-arde***	**Systematik**		
[Einheit]	Tausend$^{1-1}$	[Einheit]	[Einheit]	Million^{0}	10^{0}	[Einheit]
Tausend	Tausend^{1+0}	Tausend	Tausend	Million$^{\frac{1}{2}}$	10^{3}	Kilo
Million	Tausend^{1+1}	Million	Million	Million^{1}	10^{6}	Mega
Billion	Tausend$^{1+\mathbf{2}}$	Tausend Millionen	Milli**arde**	Million$^{1\frac{1}{2}}$	10^{9}	Giga
Trillion	Tausend$^{1+\mathbf{3}}$	Billion	**Bi**llion	Million$^{\mathbf{2}}$	10^{12}	Tera
Quadrillion	Tausend$^{1+\mathbf{4}}$	Tausend Billionen	Billi**arde**	Million$^{2\frac{1}{2}}$	10^{15}	Peta
Quintillion	Tausend$^{1+\mathbf{5}}$	Trillion	**Tri**llion	Million$^{\mathbf{3}}$	10^{18}	Exa
Sextillion	Tausend$^{1+\mathbf{6}}$	Tausend Trillionen	Trilli**arde**	Million$^{3\frac{1}{2}}$	10^{21}	Zetta
Septillion	Tausend$^{1+\mathbf{7}}$	Quadrillion	**Quadr**illion	Million$^{\mathbf{4}}$	10^{24}	Yotta

Key Ideas

- Input representation trained
- Attention to identify relevant tokens
- Two phases for training processes
- Scale up

ChatGPT

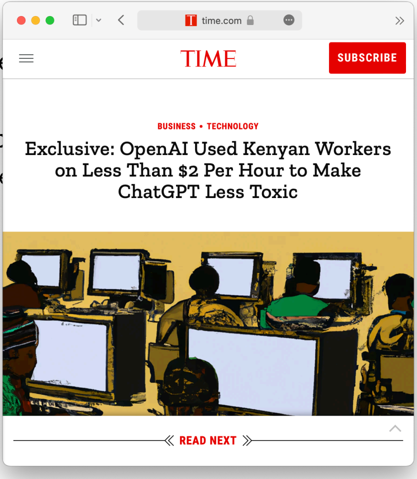
- »OpenAI« is just a name – nothing this company does is ›open‹
  - I.e., we don't know many details
- Running ChatGPT is expensive (rumors: $ 100 000 per day)
  - Usually, running a service costs a fraction of the development/training cost

ChatGPT

- »OpenAI« is just a name – nothing this company does is ›open‹
  - I.e., we don't know many details
- Running ChatGPT is expensive (rumors: $ 100 000 per day)
  - Usually, running a service costs a fraction of the development/training cost
- There are multiple ugly sides

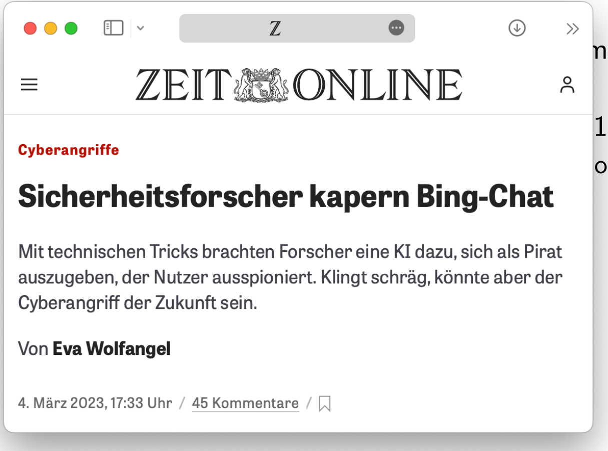
ChatGPT

- »OpenAI« is just a name – nothing this company doe
  - I.e., we don't know many details
- Running ChatGPT is expensive (rumors: $ 100 000 p
  - Usually, running a service costs a fraction of the deve
- There are multiple ugly sides



Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

ChatGPT



ZEIT ONLINE

Cyberangriffe

Sicherheitsforscher kapern Bing-Chat

Mit technischen Tricks brachten Forscher eine KI dazu, sich als Pirat auszugeben, der Nutzer ausspioniert. Klingt schräg, könnte aber der Cyberangriff der Zukunft sein.

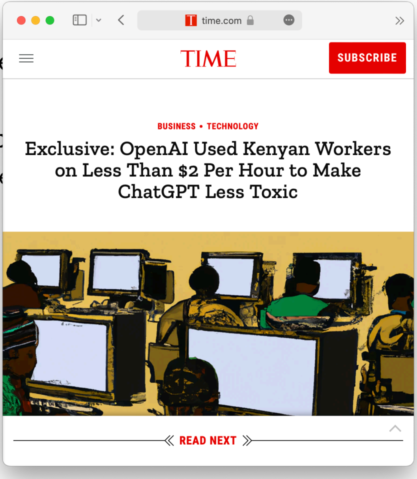
Von **Eva Wolfangel**

4. März 2023, 17:33 Uhr / 45 Kommentare

mpany doe

100 000 p

of the deve



time.com

TIME

SUBSCRIBE

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

READ NEXT

ChatGPT

- »OpenAI« is just a name – nothing this company does is ›open‹
  - I.e., we don't know many details
- Running ChatGPT is expensive (rumors: $ 100 000 per day)
  - Usually, running a service costs a fraction of the development/training cost
- There are multiple ugly sides

ChatGPT predicts probable next words

- There is no model of the world behind it
- There is no factual knowledge or reasoning about anything behind it
- No one is able to guarantee, that the produced text is factually correct

Discussion

Do we need legal regulation of »AI«, and if so, what exactly?

Section 1

Summary

Summary

- Summary