
NATURAL LANGUAGE UNDERSTANDING

VON DER FORSCHUNG IN DIE PRAXIS - DER SIEGESZUG VORTRAINIERTER SPRACHMODELLE

Sven Giesselbach – Fraunhofer IAIS
Teamleiter Natural Language Understanding

Die Fraunhofer-Gesellschaft auf einem Blick

Zahlen und Fakten



28 000
Mitarbeiterinnen
und Mitarbeiter



**74 Institute &
Forschungs-
einrichtungen**



**SEIT
1949**
seit 70 Jahren Treiber für
angewandte Forschungs-
und Entwicklung



Hauptstandorte ●
Nebenstandorte ○



Fraunhofer-Institutszentrum Schloss Birlinghoven

Forschungszentrum auf historischem Gelände



Eines der größten Forschungszentren für angewandte Informatik und Mathematik



Über 800 Wissenschaftler*innen



Kernkompetenz Künstliche Intelligenz



Ansässige Institute: IAIS, SCAI, FIT, SIT



Enge Kooperation mit regionalen Hochschulen



Hochschule
Bonn-Rhein-Sieg

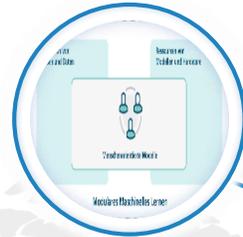


Fraunhofer IAIS

Treiber eines starken Netzwerks

ML2R

Nationales
Kompetenzzentrum für
Maschinelles Lernen



Forschungszentrum Maschinelles Lernen

Geleitet von IAIS, als Teil des Fraunhofer
Clusters of Excellence Cognitive Internet
Technologies



KI.NRW

Geschäftsstelle der
Kompetenzplattform KI in NRW



Fraunhofer-Allianz Big Data AI

Geschäftsführendes Institut der
größten Fraunhofer-Allianz mit über
30 Instituten



AI4EU

EU Leuchtturm-
Projekt für KI



International Data Spaces Association

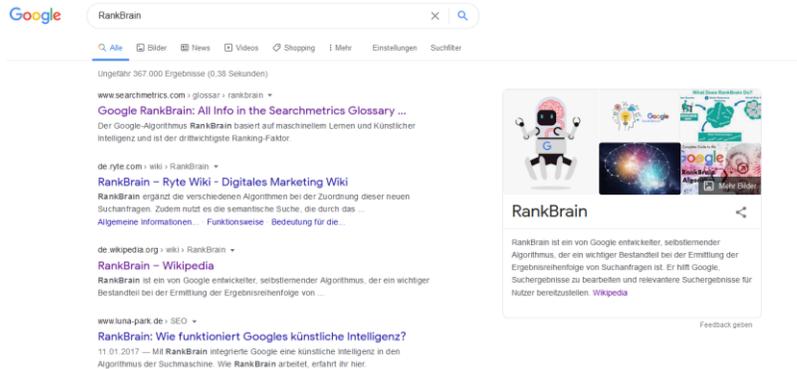
Virtueller Datenraum zum sicheren
Datenaustausch mit über 100 Unternehmen aus
unterschiedlichen Industrien

KÜNSTLICHE INTELLIGENZ

KI/NLU im Alltag

Suchmaschinen, Spracherkennung, Übersetzung, Sprachassistenten

Intelligente Suchmaschinenergebnisse



www.google.de

Live-Transkription

TRANSKRIPTION

Skype schreibt künftig Gespräche mit

Skype bietet künftig eine Transkription von Gesprächen an, damit gehörlose oder schwerhörige Menschen Audio- und Videokonferenzen folgen können. 2019 sollen auch bei Powerpoint Echtzeituntertitel eingeführt werden, die bei der Präsentation eingeblendet werden.

4. Dezember 2018, 7:34 Uhr, Andreas Donath



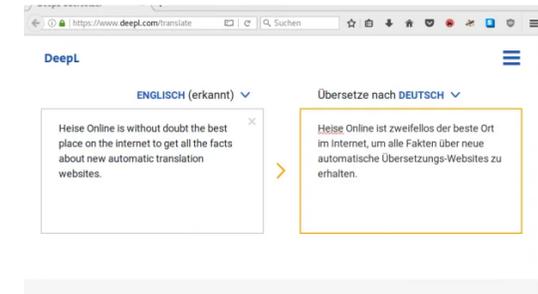
<https://www.golem.de/news/transkription-skype-schreibt-kuenftig-gespraechе-mit-1812-138047.html>

Sprachübersetzung

Maschinelle Übersetzer: DeepL macht Google Translate Konkurrenz

Linguae heißt jetzt DeepL und geht mit einem neuen maschinellen Übersetzer auf den Markt. Der basiert auf tiefen Convolutional Networks und dürfte effizienter übersetzen als Google Translate, das rekurrente Netze einsetzt.

Lesedzeit: 2 Min. In Pocket speichern



An dieser Übersetzung haben wir nichts auszusetzen.

29.08.2017 07:31 Uhr

Von Pina Merkert

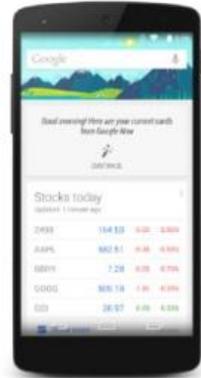
<https://www.heise.de/newsticker/meldung/Maschinelle-Uebersetzer-DeepL-macht-Google-Translate-Konkurrenz-3813882.html>

KI/NLU im Alltag - II

Suchmaschinen, Spracherkennung, Übersetzung, Sprachassistenten



Apple Siri (2011)



Google Now (2012)



Microsoft Cortana (2014)



Amazon Alexa/Echo (2014)



Facebook M & Bot (2015)



Google Home (2016)



Apple HomePod (2017)



Samsung Bixby (2017)

KI/NLU in den Schlagzeilen

GPT-3 und DALL-E

DIGITALE REVOLUTION

Dieses Sprachprogramm verblüfft Experten – und birgt Potenzial für die Wirtschaft

Die KI-Software GPT-3 schreibt Texte auf erstaunlich hohem Niveau – obwohl sie die Inhalte nicht versteht. Die Anwendungsmöglichkeiten sind vielfältig.

<https://www.handelsblatt.com/technik/digitale-revolution/digitale-revolution-dieses-sprachprogramm-verbluefft-experten-und-birgt-potenzial-fuer-die-wirtschaft/26126390.html>

● This article is more than 5 months old

A robot wrote this entire article. Are you scared yet, human?

GPT-3

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

● For more about GPT-3 and how this essay was written and edited, please read our editor's note below

<https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>

Machine Learning: KI GPT-3 erstellt unter dem Künstlernamen DALL-E Bilder

Eine spezielle Version des von OpenAI entwickelten Sprachmodells GPT-3 erstellt Bilder anhand von Beschreibungen.

Lesezeit: 2 Min.  In Pocket speichern

   13



(Bild: agsandrew / Shutterstock.com)

<https://www.heise.de/news/Machine-Learning-GPT-3-erstellt-unter-dem-Kuenstlernamen-DALL-E-Bilder-5005847.html>

Was ist (hybride) künstliche Intelligenz?

Eine (von vielen möglichen) Definitionen

Machine Learning

- z.B. Neuronale Netze

Rechenpower

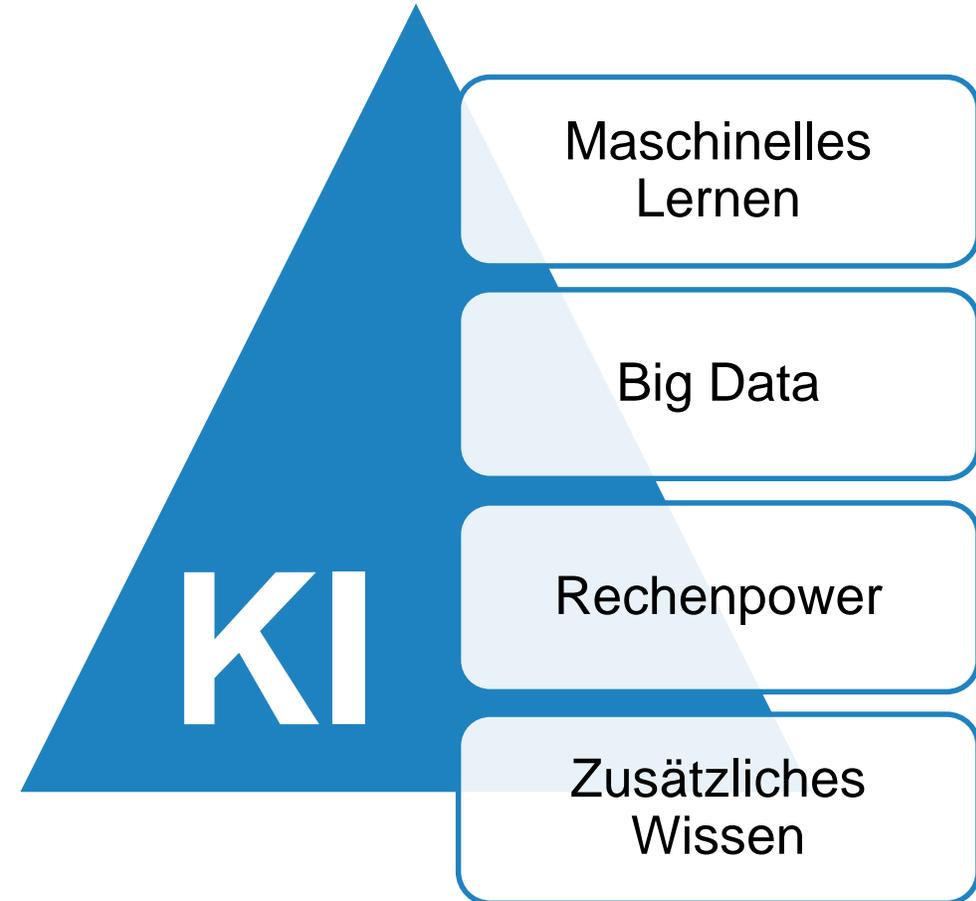
- Wachsende Computer-Power und immer günstiger werdende Technologien

Big Data

- z.B. Posts aus sozialen Medien, Sensordaten etc.

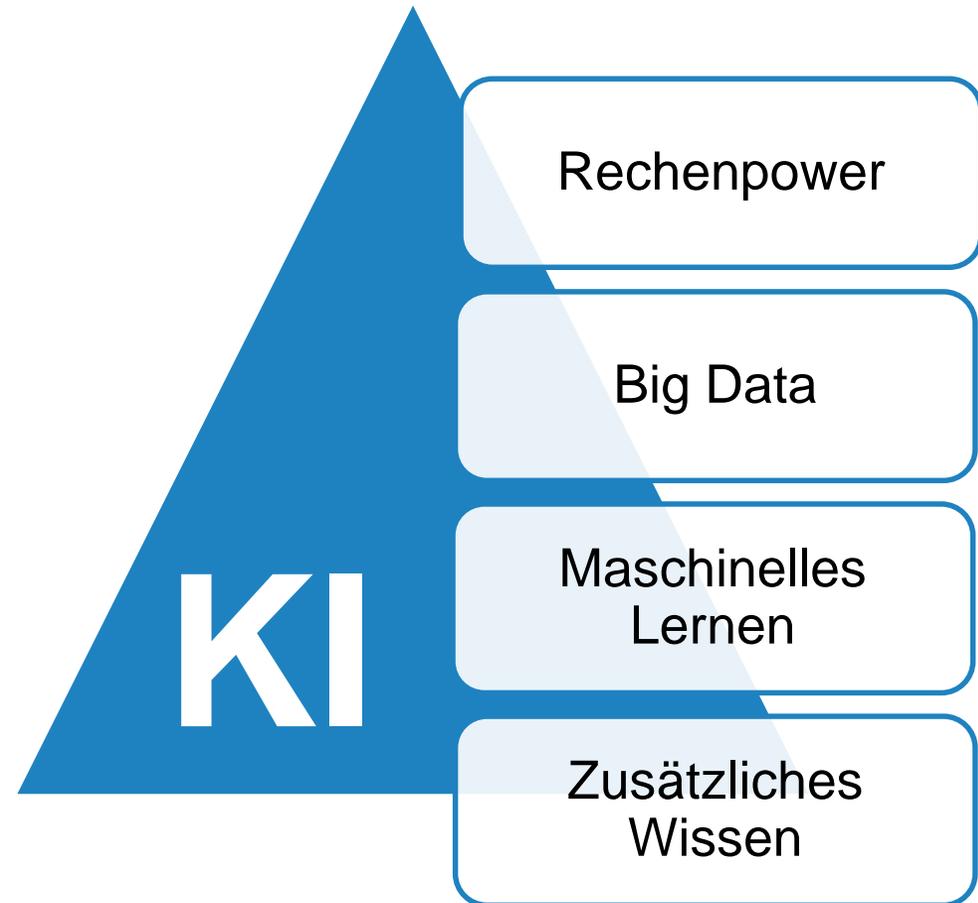
Integration von Wissen

- z.B. über Regeln, Ontologien, Probabilistic Soft Logic, etc.



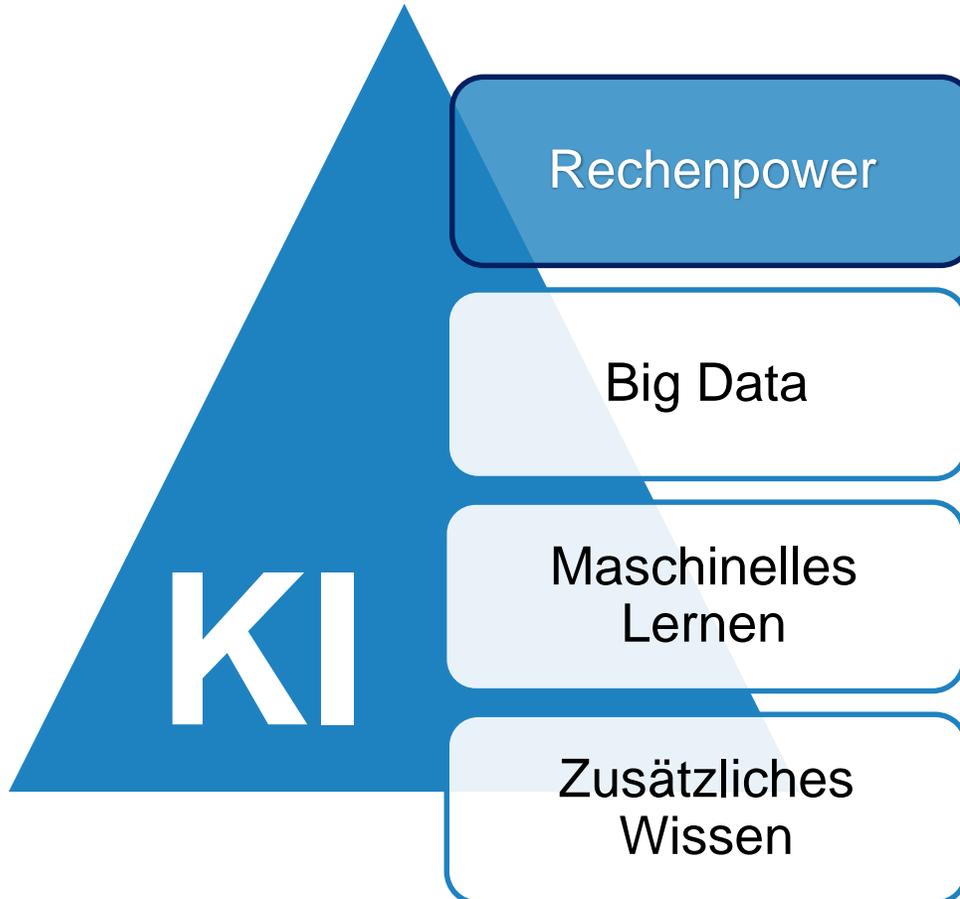
Komponenten Künstlicher Intelligenz

Beispiel GPT-3



Komponenten Künstlicher Intelligenz

Beispiel GPT-3 - Rechenpower



We are waiting for OpenAI to reveal more details about the training infrastructure and model implementation. But to put things into perspective, GPT-3 175B model **required 3.14E23 FLOPS of computing for training**. Even at theoretical [28 TFLOPS](#) for V100 and lowest 3 year reserved cloud pricing we could find, this will take **355 GPU-years and cost \$4.6M for a single training run**. Similarly, a single RTX 8000, assuming 15 TFLOPS, would take 665 years to run.

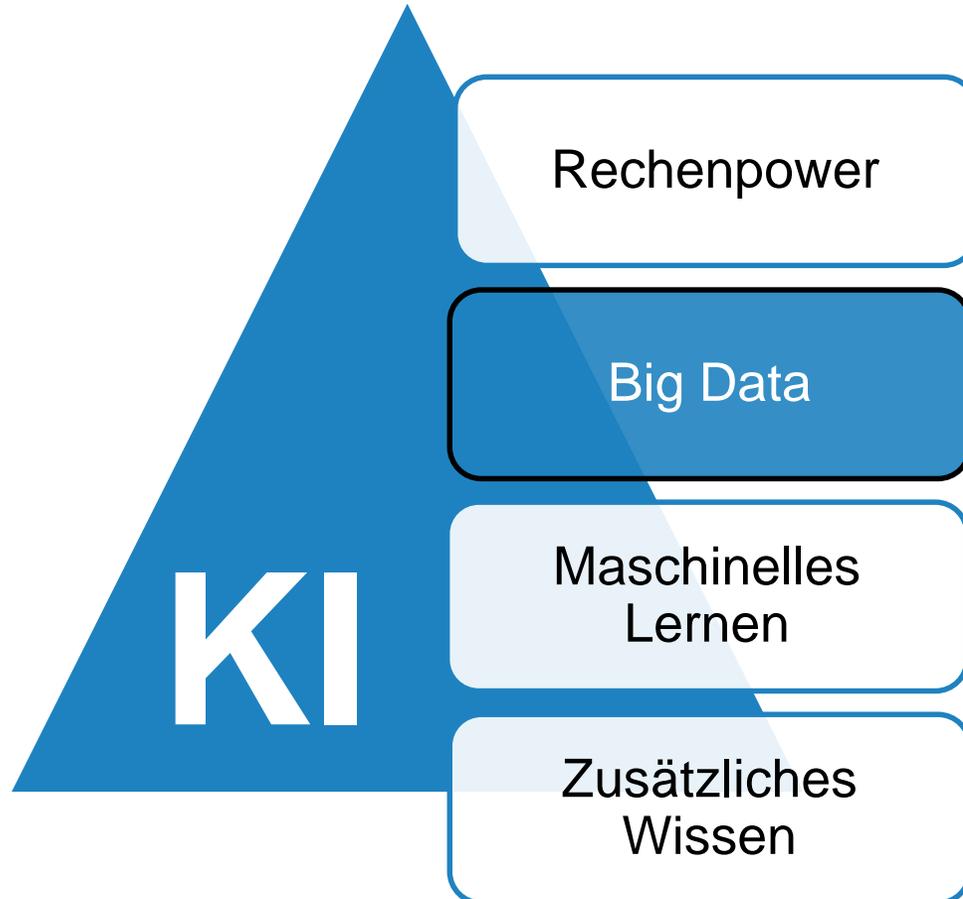
<https://lambdalabs.com/blog/demystifying-gpt-3/>

“The supercomputer developed for OpenAI is a single system with more than **285,000 CPU cores, 10,000 GPUs and 400 gigabits per second** of network connectivity for each GPU server,” the companies stated in a [blog](#).

<https://news.developer.nvidia.com/openai-presents-gpt-3-a-175-billion-parameters-language-model/>

Komponenten Künstlicher Intelligenz

Beispiel GPT-3 – Big Data



■ GPT-3 wurde auf **570 GB Texten** trainiert

■ Anzahl Dokumente pro Sprache:

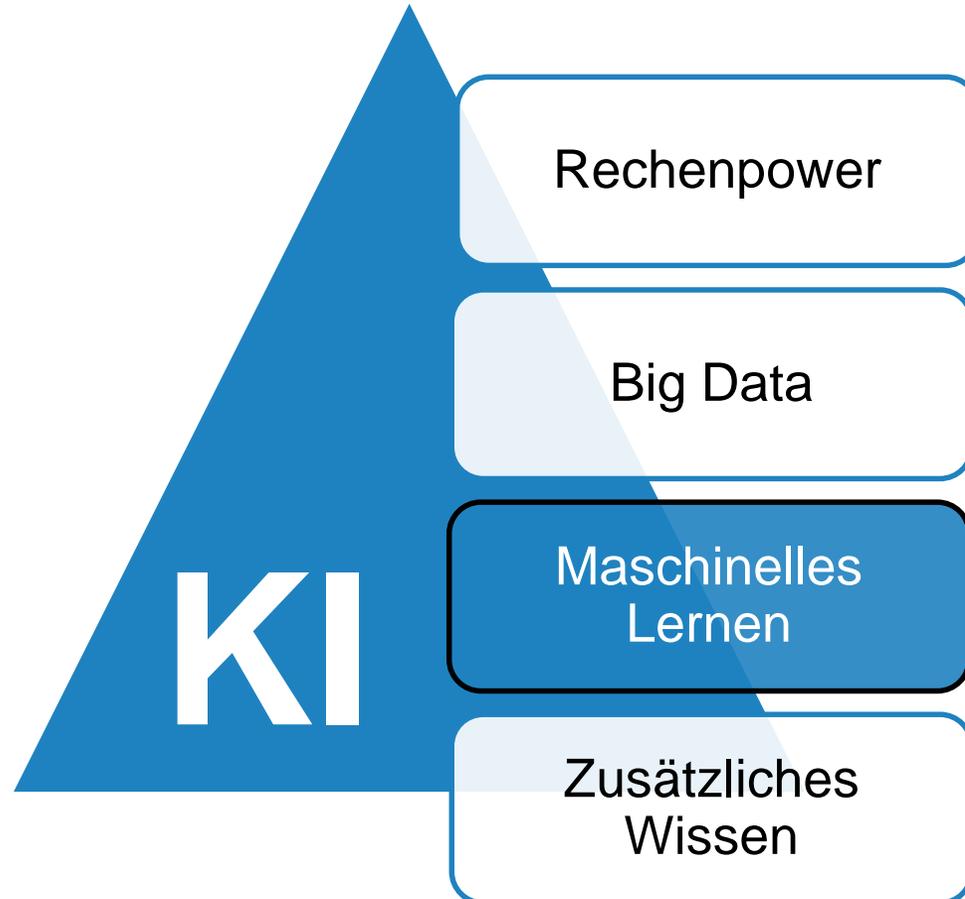
Englisch	235.987.420
Deutsch	3.014.597
Französisch	2.568.341

■ Paper zeigt, dass die Qualität des Modells von der **Größe des Modells, der Anzahl der Daten und der Trainingsdurchläufe** abhängt

■ Insgesamt beinhalten die Texte ca. **500 Milliarden Token** (Achtung: 1 Token != 1 Wort)

Komponenten Künstlicher Intelligenz

Beispiel GPT-3 – Maschinelles Lernen



- GPT-3 basiert auf dem „**Decoder**“ der neuronalen „**Transformer**“-Architektur
 - Die größte Variante ist ein sehr **tiefes neuronales Netz** mit **96 „Attention“ Schichten** und insgesamt **175 Milliarden Parametern**
- Das Original Transformer-Paper „**Attention is all you need**“ (Vaswani et al., 2017) hat heute bereits ca. **17.700 Zitierungen**

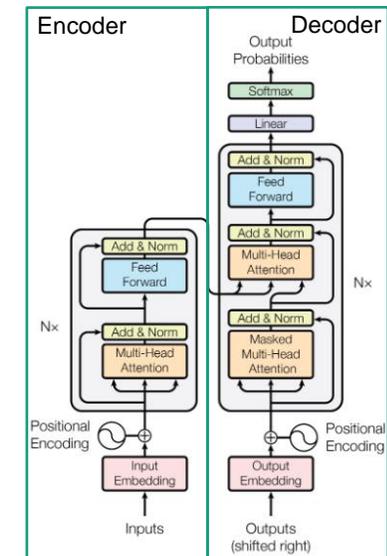
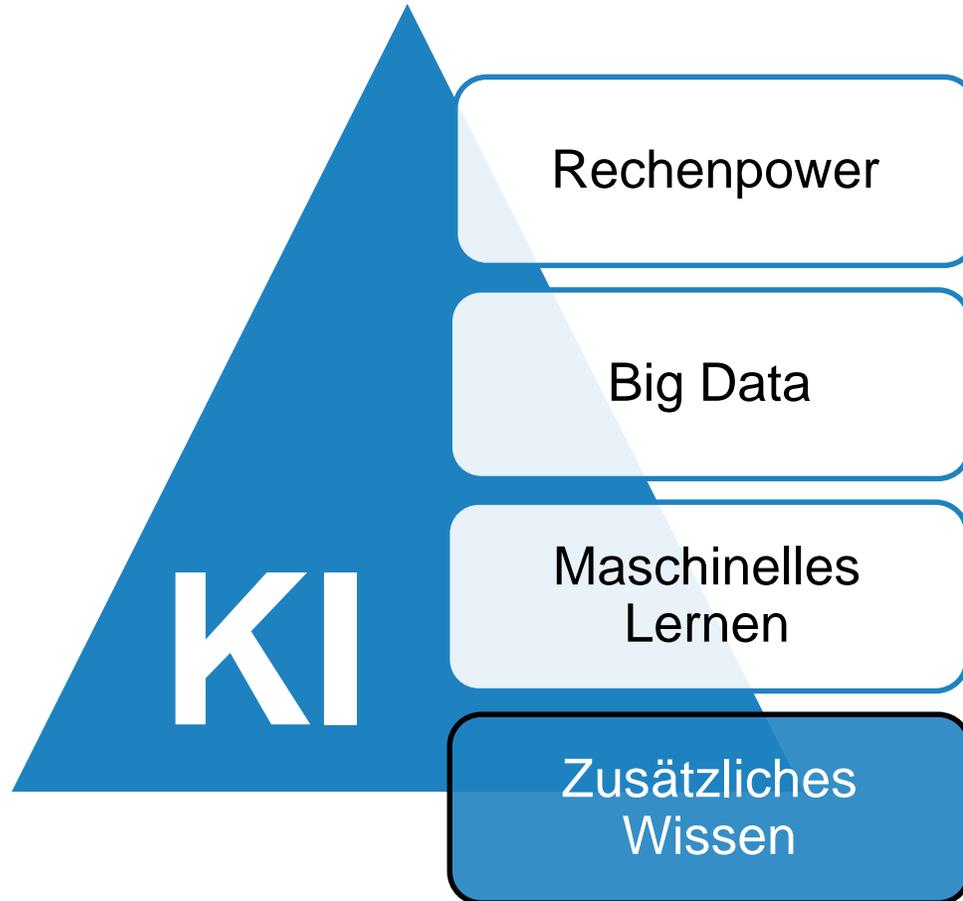


Figure 1: The Transformer - model architecture.

Vaswani et al. – Attention is all you need (2017)

Komponenten Künstlicher Intelligenz

Beispiel GPT-3 – Zusätzliches Wissen



- GPT-3 nutzt **kein Wissen** außerhalb der Trainingstexte die es sieht, dies lässt **Potenzial** offen:
 - „[...] it still **sees** much **more text** [...] **than a human** sees in **their lifetime** [...]” – (Brown et al. 2020 – Language Models are Few Shot Learners)
 - “[...] apparently simple problems require humans to **integrate knowledge across vastly disparate sources** [...] entirely **different sorts of tools** are needed, along with deep learning, if we are to **reach human-level cognitive flexibility**.” – (Gary Marcus 2018 – Deep Learning – A Critical Appraisal”)

NATURAL LANGUAGE UNDERSTANDING

Natural Language Understanding

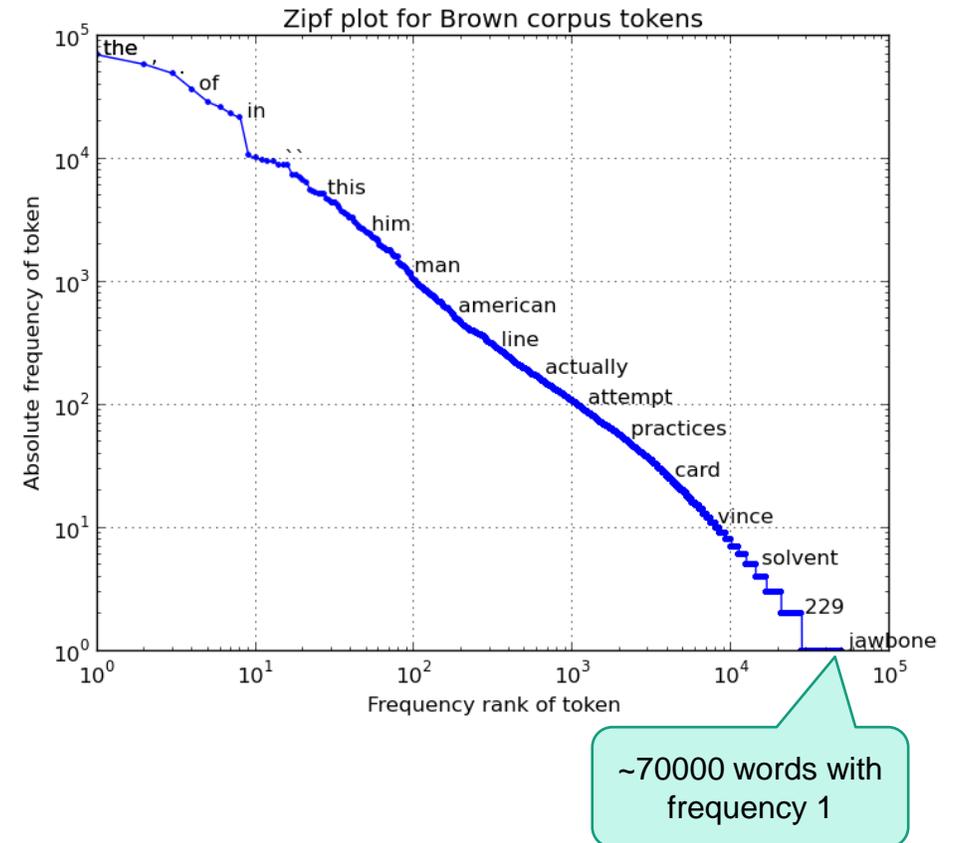
Einordnung und Ziele

- **Natural Language Understanding (NLU)** ist ein Teil von **Natural Language Processing (NLP)** und befasst sich mit dem **Verstehen natürlicher Sprache**. Es liegt an der Schnittstelle von:
 - Informatik,
 - Künstlicher Intelligenz,
 - und Linguistik
- **Ziel** ist es, Computern die Fähigkeit zu geben, Sprache zu verarbeiten bzw. zu verstehen um **nützliche Aufgaben zu erledigen**, wie z.B.:
 - Automatische Dokumentenverarbeitung
 - Fragen beantworten
 - Siri, Google Assistant, Facebook Alexa, Cortana ...
- Natürliche Sprache vollständig verstehen und abzubilden ist sehr schwer
 - Die Aufgabe gilt als **AI-Complete**
 - d.h. Computer brauchen dafür menschenähnliche Intelligenz

Was ist besonders an NLP/NLU?

Unterschiede zu anderen Feldern

- Große Anzahl an “Features” ($\gg 100.000$), die genutzt werden können um Dokumente darzustellen
- Lange Abhängigkeiten im Input müssen beachtet werden um Dokumente richtig zu verstehen
- Gerade seltene Features und Muster können besonders bedeutsam sein
- Eine riesige Anzahl bedeutsamer Muster
- Eine riesige Menge potentieller Daten
 - Dokumente in (Firmen-)Datenbanken
 - Das Internet



Maschinelles Lernen auf Text

Vektorraum-Repräsentationen

- Um maschinelles Lernen auf textuellen Daten zu nutzen, benötigen wir **numerische Repräsentationen von Text**
- Eine einfache Repräsentation für **Dokumente** wäre ein Vektor v für den gilt
 - v hat für jedes Wort in unserem **Vokabular** einen Eintrag
 - Jeder Eintrag reflektiert die **Häufigkeit** des jeweiligen Wortes im Dokument
 - So eine Repräsentation nennt sich **Bag-of-Words Repräsentation**
 - **Idee: Ähnliche Dokumente** erhalten ähnliche **Repräsentationen**

Die Maus isst die Käsestücke

Er legt die Maus auf die Matte

käsestücke:1 isst:1 er:0 maus:1 auf:0 legt:0 matte:0 die:2

käsestücke:0 isst:0 er:1 maus:1 auf:1 legt:1 matte:1 die:2

- Es gibt einige **Probleme**, unter Anderem:
 - Was ist mit **Synonymen**?
 - Was ist mit **Homonymen**?

Anwendung im Bereich Xtreme MultiLabel Classification

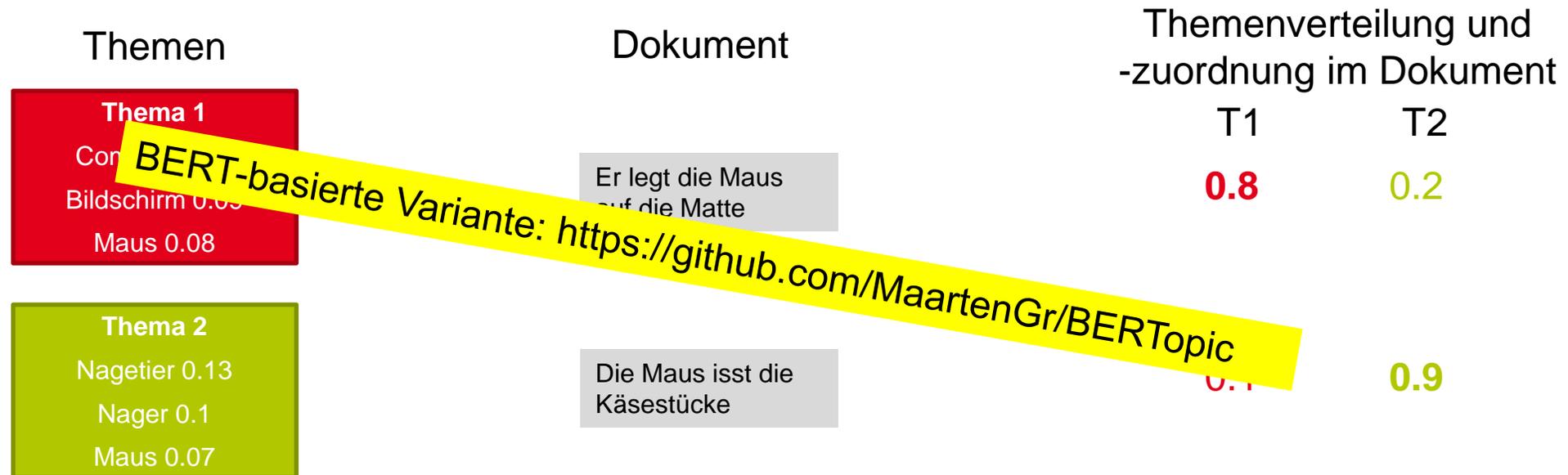
- BoW (bzw. Tf-idf) Features werden häufig z.B. für Xtreme Mutlilabel Classification Modelle genutzt*
 - Aufgaben bei denen verschiedene Label einem Dokument zugeordnet werden müssen
 - Meistens aus einer riesigen Menge Labels $|Y| \gg 100$
- Beispielprojekte:
 - CPC Klassifikation von Patenten
 - Taggen von Nachrichtenartikeln
 - Klassifikation von Produkten

*Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising - Prabhu et. al - 2018

Maschinelles Lernen auf Text

Topic Models

- Eine Repräsentationen durch **unüberwachtes Lernen** bieten so genannte **Topic Models**
 - Das bekannteste Beispiel dafür ist die **Latent Dirichlet Allocation (LDA)** von Blei et al. (2003)
 - LDA lernt automatisch **Themencuster** durch das Analysieren **großer Textmengen**
 - **Dokumente** werden als **Verteilung** von **Themen** repräsentiert
 - **Themen** sind **Verteilungen** der **Wörter** die in Ihnen vorkommen

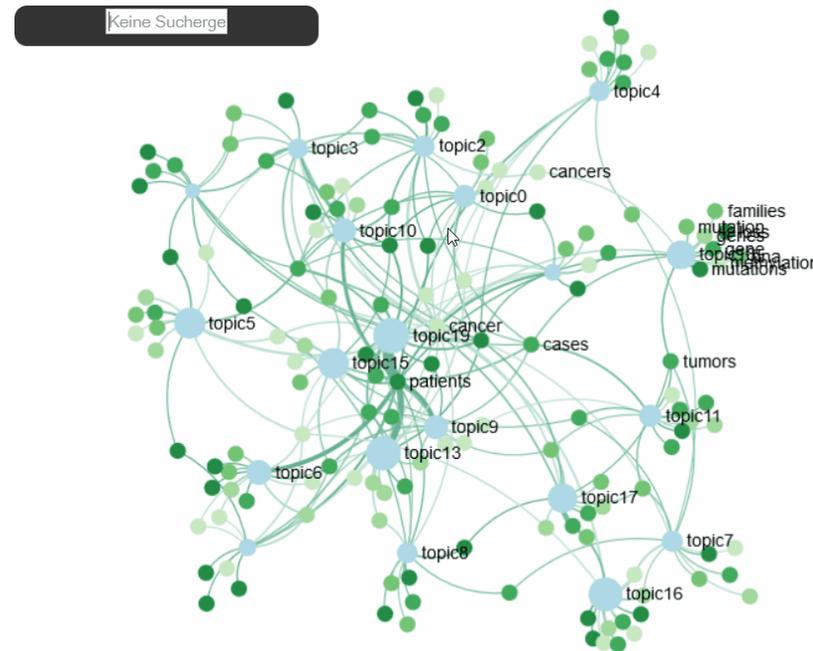


Anwendungsbeispiel: Interaktives Inhaltsverzeichnis von Dokumentenkollektionen

- Blaue Knoten: Themen
- Grüne Knoten: Wichtigste Themenwörter

- Beispiel trainiert auf ~200.000 PubMed abstracts

- Anwendungen in:
 - Social Media Analyse
 - Analyse von Patenten
 - Analyse von Publikationen



topic18
digest

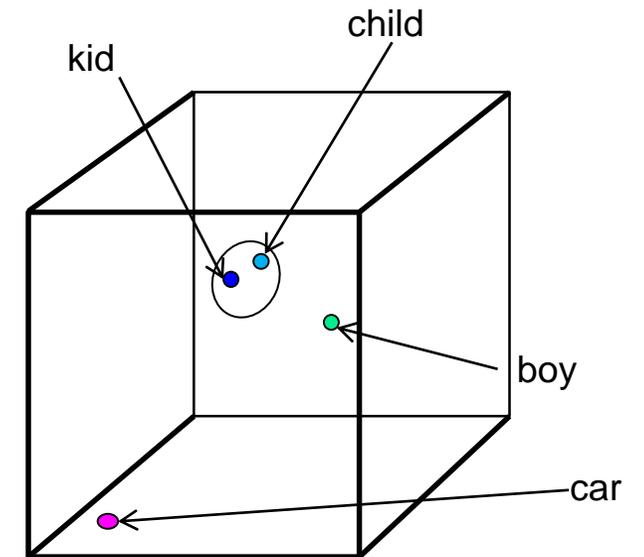
Lynch syndrome is the most common inherited colon cancer syndrome. Patients with Lynch syndrome develop a range of cancers including colorectal cancer (CRC) and carry a mutation on one of the mismatched repair (MMR) genes. Although CRC usually occurs after the fourth decade in patients with Lynch syndrome harboring a heterozygous MMR gene mutation it can occur in children with Lynch syndrome who have a compound heterozygous or homozygous MMR gene mutation. We report a case of CRC in a 13-year-old patient with Lynch syndrome and congenital heart disease. This patient had a heterozygous mutation in MLH1 (an MMR gene) but no compound MMR gene defects and a K-RAS somatic mutation in the cancer cells.

BACKGROUND: Changes in the number of alleles of a chromosome may have an impact upon gene expression. Loss of heterozygosity (LOH) indicates that one allele of a gene has been lost and knowing the exact copy number of the gene would indicate whether duplication of the remaining allele has occurred. We

Maschinelles Lernen auf Text

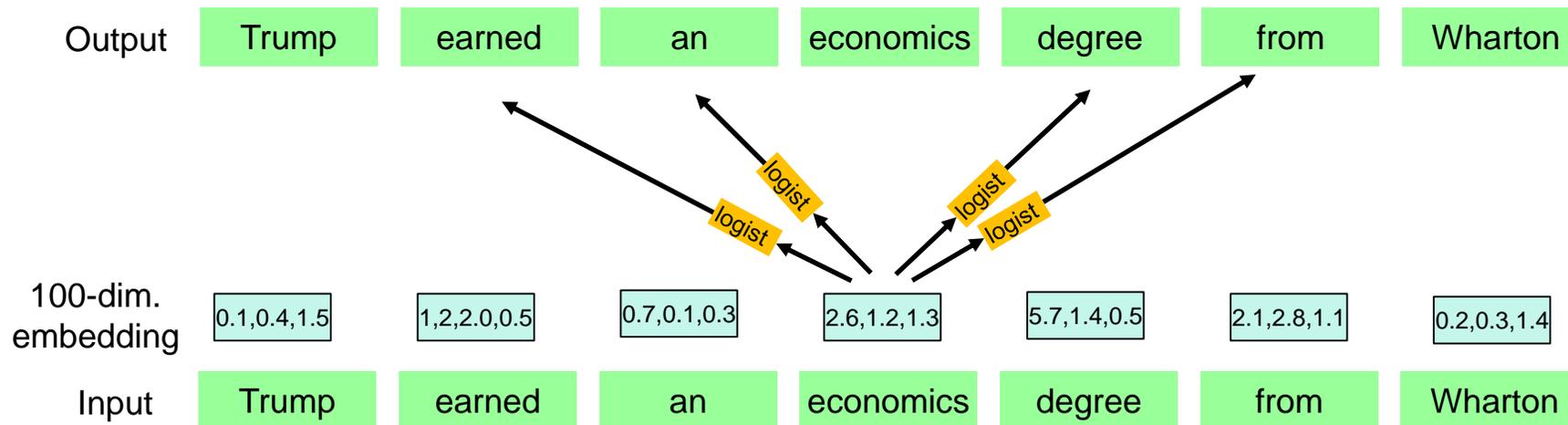
Embedding-Repräsentationen I

- Hypothese: **Wörter** erhalten ihre **Bedeutung** aus den Wörtern in ihrem **Kontext**
- Ziel ist es eine Repräsentation zu erhalten, so dass:
 - Wörter **ähnlicher Bedeutung** → **ähnliche Vektoren** erhalten
[Wikipedia: Distributional semantics]
 - **Unähnliche** Wörter, weit **voneinander entfernt** sind



Word2Vec [Mikolov et al. 2013]

- Jedes Wort bekommt ein Embedding der Größe k zugewiesen, z.B. $k = 100$
- Sage die **benachbarten** Wörter über das Embedding des zentralen Wortes vorher
- Vorhersage über ein einfaches Modell: linear logistic model



- Kann einfach für ungesehene Wörter (Fasttext) und Dokumente (Doc2Vec) erweitert werden
[Bojanowski et al. 2017] [Le & Mikolov 2014]

Maschinelles Lernen auf Text

Selbstüberwachtes Lernen

- Eine **Mischform** aus **überwachtem** und **unüberwachtem** Lernen stellt das selbstüberwachte Lernen dar
 - Eine überwachte Lernaufgabe, wird erstellt in dem **Teile der Trainingsdaten ausgelassen** und **wieder vorhergesagt** werden müssen
- Selbstüberwachtes Lernen erfordert **keine Annotationen**
- Diese Form des Lernens ist heute ein Grundbaustein aller Text-Repräsentationen
 - Durch selbstüberwachtes Lernen auf großen Textbeständen bauen Modelle ein Verständnis von Wortbedeutungen auf
 - Diese Modelle können dann mit vergleichsweise wenig annotierten Trainingsdaten auf weiterführende Aufgaben spezialisiert werden

Anwendung: Aufbau von Terminologien

- Gegeben einem Eingabewort, liefere Kandidaten für Synonyme und verwandte Wörter
- Vorschläge können in eine Terminologie überführt werden

Wort

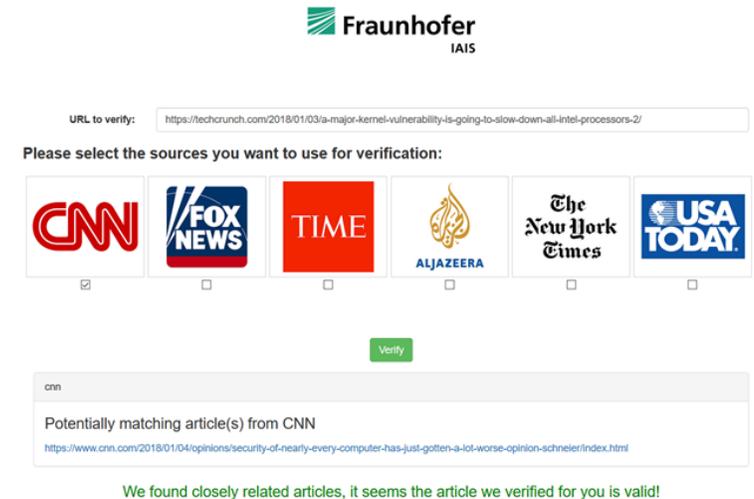
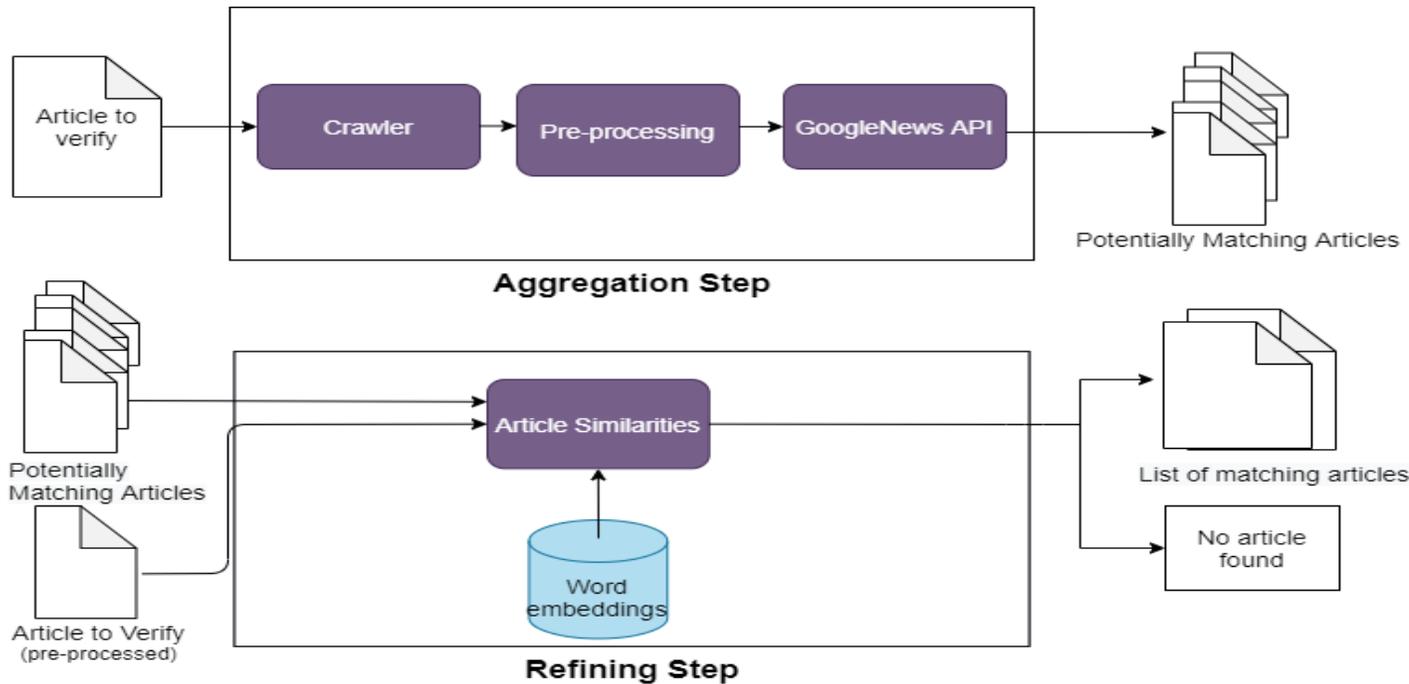
Tumor

Ausführen

w2v	glove	fasttext
Muskel 0.76	Neuroendokriner 0.61	Tumors 0.9
Lymphknoten 0.76	Tumoren 0.6	Tumore 0.89
Abszess 0.74	Tumors 0.54	Lymphom 0.89
Gendefekt 0.73	Metastasen 0.54	Karzinom 0.88
Schilddrüse 0.73	neuroendokriner 0.53	maligner_Tumor 0.88

Anwendung: Document Retrieval für Nachrichtenvalidierung

- Gegeben einem Ausgangsartikel, berechne die Ähnlichkeit zu Artikeln von bekannten Quellen



Supporting verification of news articles with automated search for semantically similar articles – Gupta et al. - ROMCIR 2021

Distributed Embeddings vs. Contextualized Embeddings

- Embedding Methoden wie **Word2Vec** und **FastText** sind bekannt als „**distributed embeddings**“
 - Sie lernen ein **einzelnes Embedding pro Wort** über all seine Kontexte hinweg
 - Gut geeignet für Abfragen nach **nächsten Nachbarn**
 - **Nicht ideal für weiterführende Aufgaben**
- Eine „neuere“ Art von Embeddings sind „**contextualized**“ Embeddings
 - Sie lernen Wortembeddings **pro Kontext** in dem ein **Wort** auftaucht
 - Sie schneiden die **Embeddings** auf diesen **Kontext zu** (z.B. nützlich bei **Homonymen**)

Maschinelles Lernen auf Text

Embedding-Repräsentationen II

- Die heute prominentesten Modelle zur Erzeugung von Text-Repräsentationen sind BERT und GPT
- BERT (Devlin et al., 2018)
 - Basiert auf dem “Encoder” des Transformer Netzwerkes und lernt Repräsentationen in dem es fehlende Wörter vorhersagt
- GPT (Jaolin et al., 2019)
 - Basiert auf dem “Decoder” des Transformer Netzwerkes und lernt Repräsentationen in dem es die nächsten Wörter eines Satzes vorhersagt

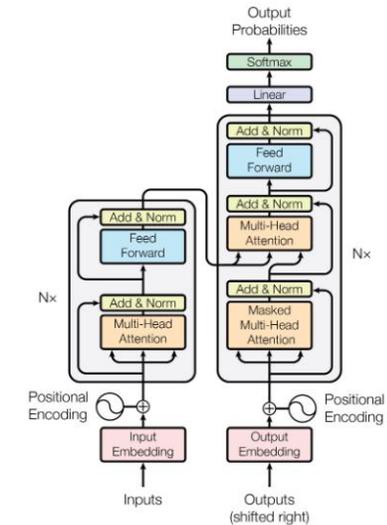


Figure 1: The Transformer - model architecture.

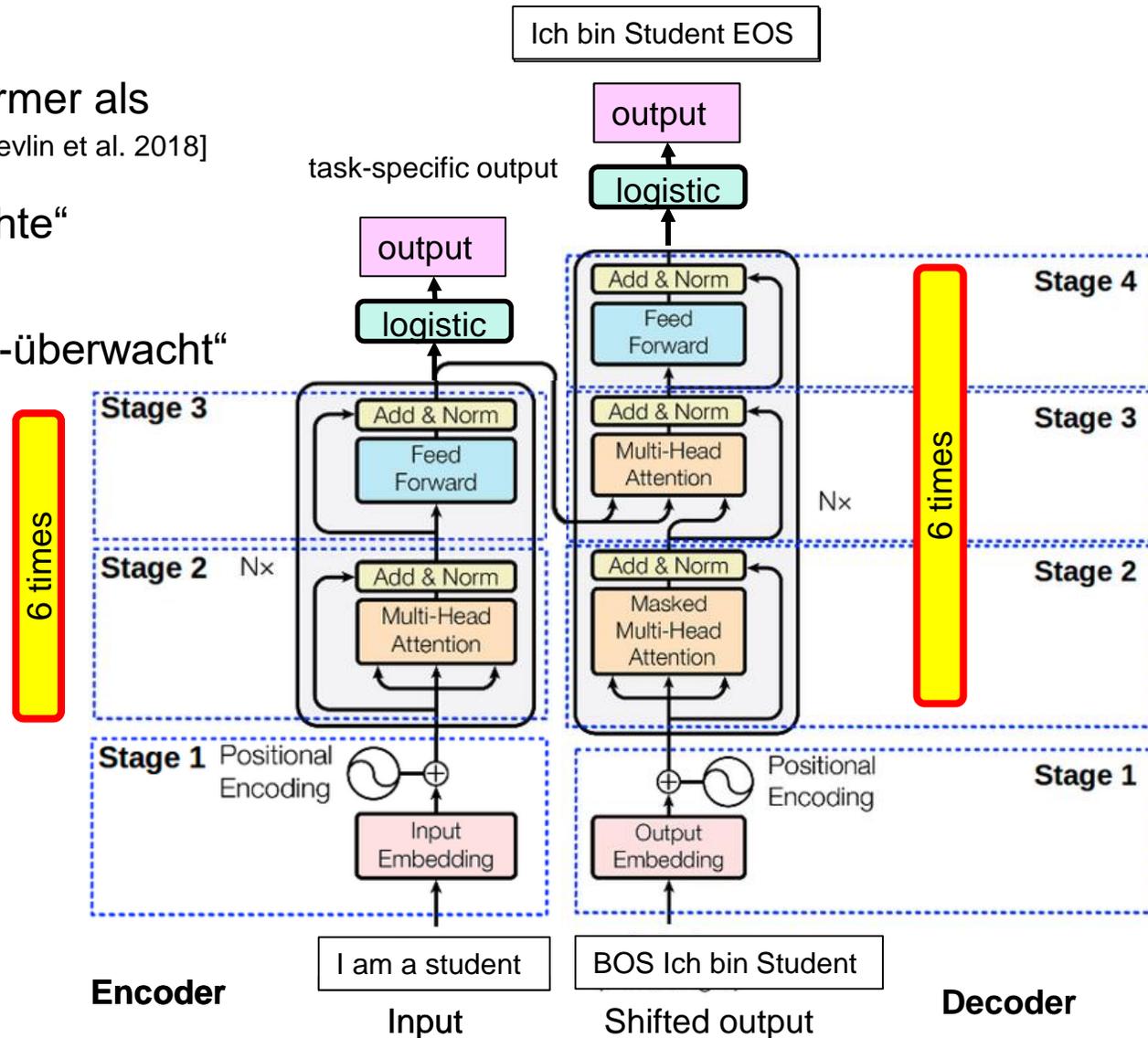
Vaswani et al. – Attention is all you need (2017)

BERT: Transformer Encoder

- Nutzt die Encoder-layer des Transformer als Sprachmodell [Devlin et al. 2018]
- Nutzt eine spezielle „selbst-überwachte“ Lernaufgabe
- Wird auf großen Textmengen „selbst-überwacht“ trainiert

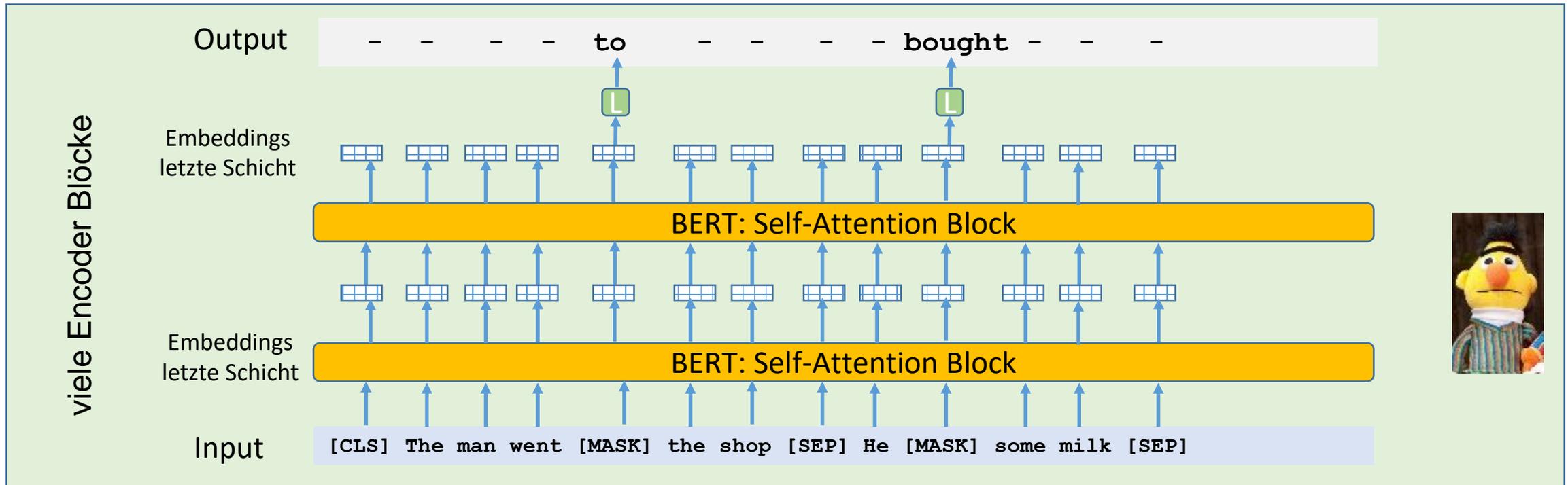


BERT
Bidirectional Representations
from Transformers

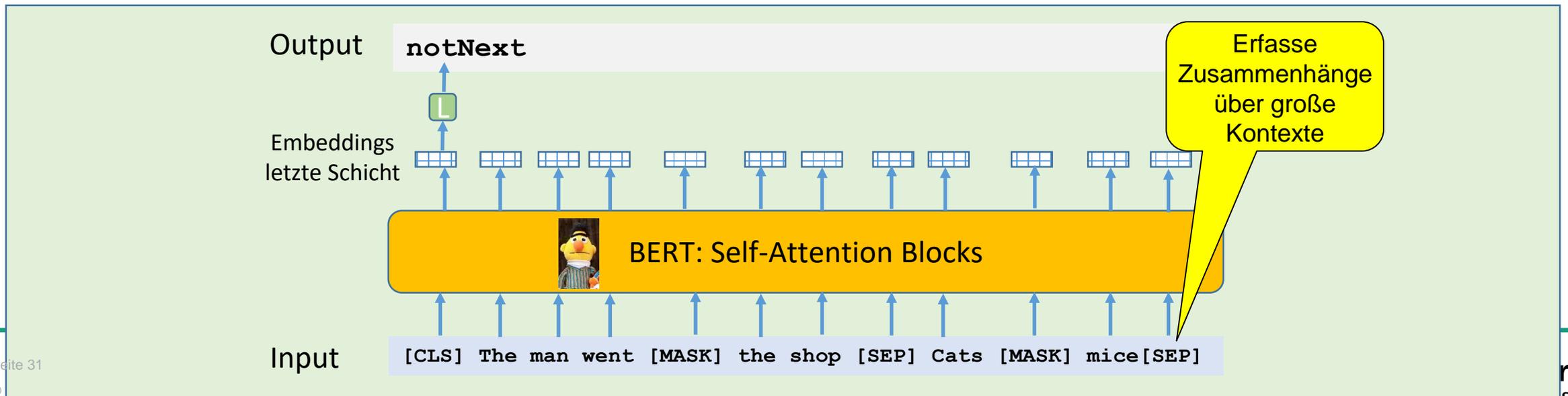
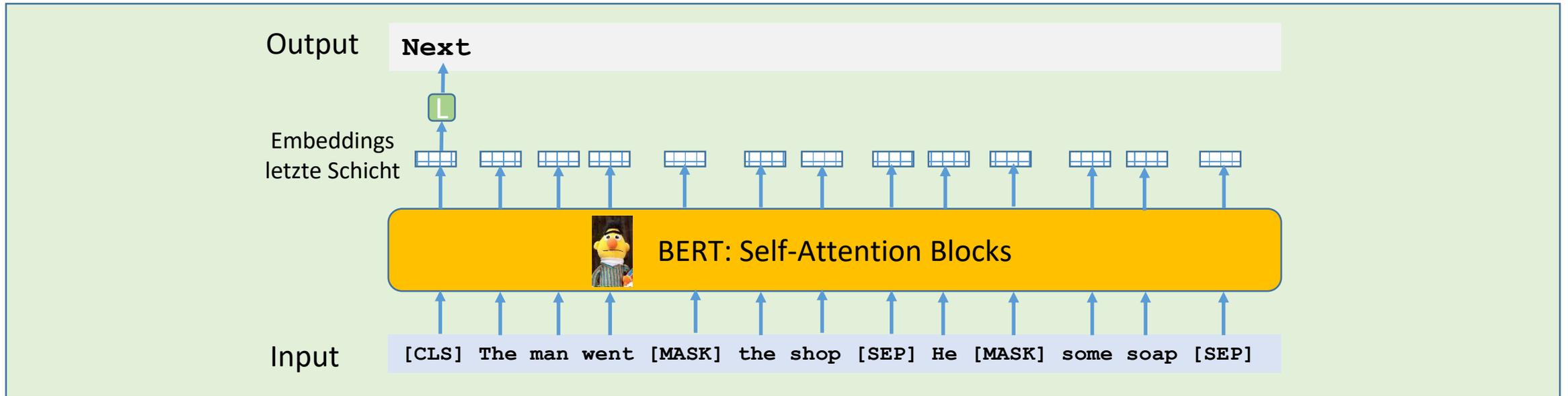


Vortrainings-Aufgabe 1: Maskierte Wörter vorhersagen

- Start token [CLS]. Ende ersten und zweiten Text mit [SEP]
- Ersetze zufällig 15% der Wörter mit [MASK]
- Sage diese Wörter mittels logistischem Klassifizierer in der letzten Schicht vorher



Vortrainings-Aufgabe 2: Aufeinanderfolgende Sätze erkennen



BERT: Vortrainierte Modelle

- Benötigt einen großen Korpus für das Vortrainieren
 - BooksCorpus 800M Wörter
 - Englische Wikipedia 2500M Wörter

	# Schichten	hidden dim	# param	Trainingszeit
BERT base	12	768	110M	4 days on 4 TPUs
BERT large	24	1024	340M	4 days on 16 TPUs

BERT: Finetuning

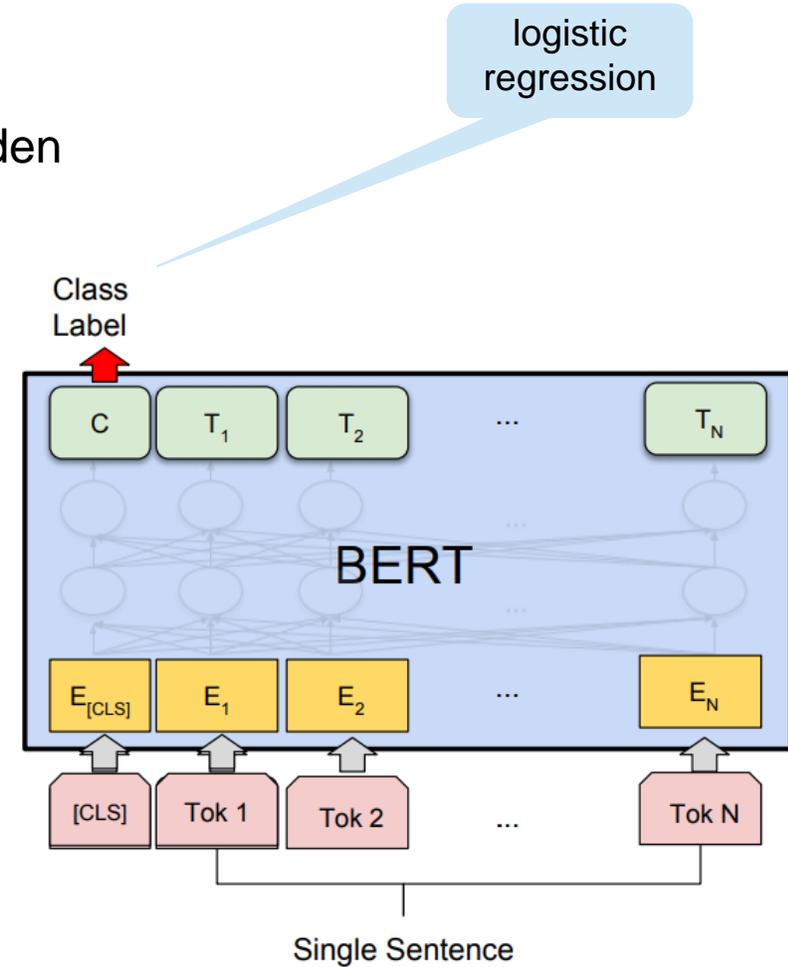
- Vortrainierte Modelle verstehen die Struktur der Sprache
 - Syntax: Wie Wörter einen Satz bilden können
 - Semantik: Wie Fakten oder Relationen ausgedrückt werden
- Idee: Adaptiere solche Modelle auf einen neuen Task
Beispiel: Klassifiziere die Meinung in einem Satz:
positiv, negativ, neutral

■ Yet the act is still charming here.

positiv

■ This isn't a new idea.

negativ



Anwendungsbeispiel: Relationsextraktion

- Erkenne die Relationen zwischen verschiedenen Entitäten
 - Verknüpfe sie wenn sie in einer Relation stehen
 - Erkenne die Klasse ihrer Relation

Anmerkungen auswählen

Steve Jobs , damals CEO von Apple , trug wie immer einen Rollkragenpullover.

ENTITY1 ENTITY2

Löschen

Ausführen

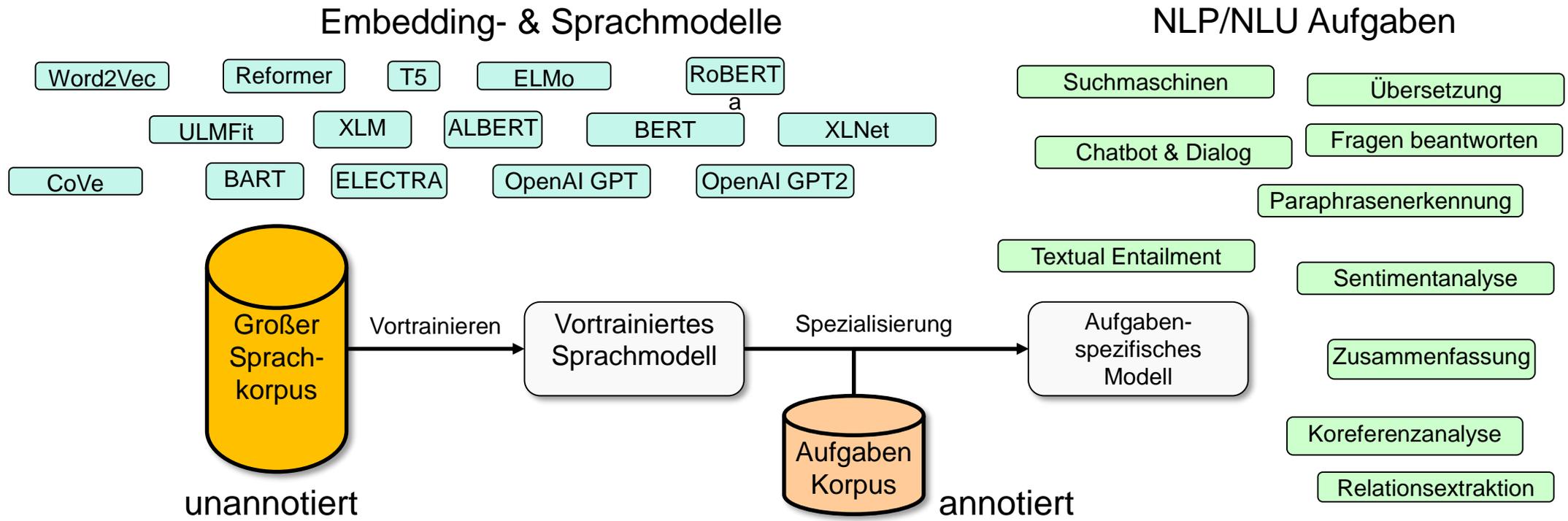
entity1 --OrganizationLeadership--> entity2

Steve Jobs, damals CEO von Apple, trug wie immer einen Rollkragenpullover.

Maschinelles Lernen auf Text

Das neue Paradigma von NLP und NLU

- Lerne ein **Sprachmodell** auf einem **großen** unannotierten Sprachkorpus (**Vortrainieren**)
- Erhalte aus dem Sprachmodell **kontextsensitive Repräsentationen**
- Ergänze das Modell (z.B. mit einer Klassifikationsschicht) und trainiere es auf **annotierten** Daten für eine **bestimmte Aufgabe** (**Spezialisierung**)



Maschinelles Lernen auf Text

Verfügbare Ressourcen

Datensätze:

- **Huggingface datasets**
<https://huggingface.co/docs/datasets/>
- **Kaggle datasets**
<https://www.kaggle.com/datasets?fileType=csv>
- **TensorFlow datasets**
<https://www.tensorflow.org/datasets>
- **Google Dataset Search**
<https://datasetsearch.research.google.com/>

Vortrainierte Transformer-Modelle:

- **Huggingface Model Hub**
<https://huggingface.co/models>

13:45 Uhr

Lydia Pintscher • Wikimedia Deutschland e.V.

▼ Wikidata : Datenqualität in einer offenen Wissensdatenbank

15:15 Uhr

Stefan Geißler • Kairntech

► Open Source - Ein Schlaraffenland für KI/NLP-Anwendungen

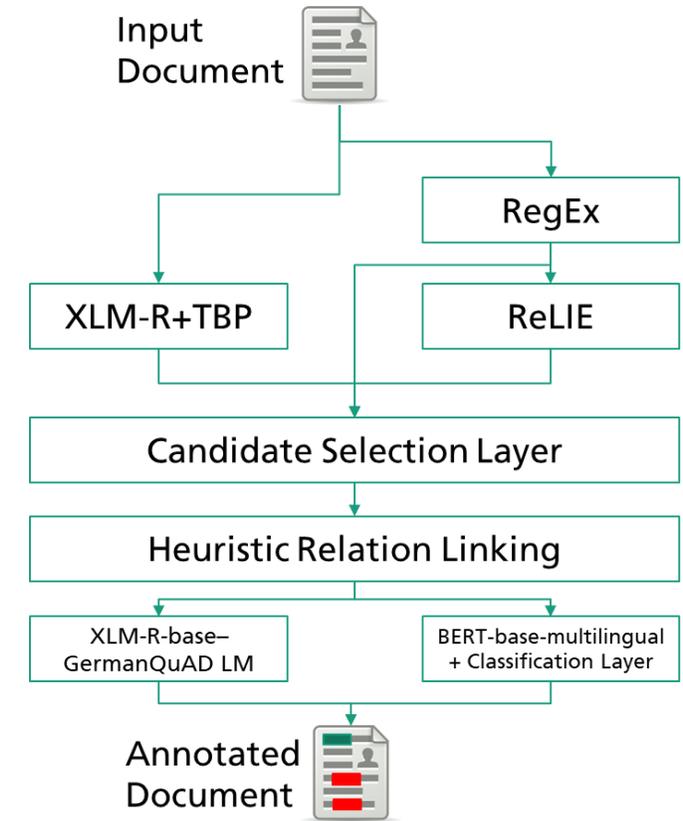
Wie arbeiten wir mit Transformer-Modellen?

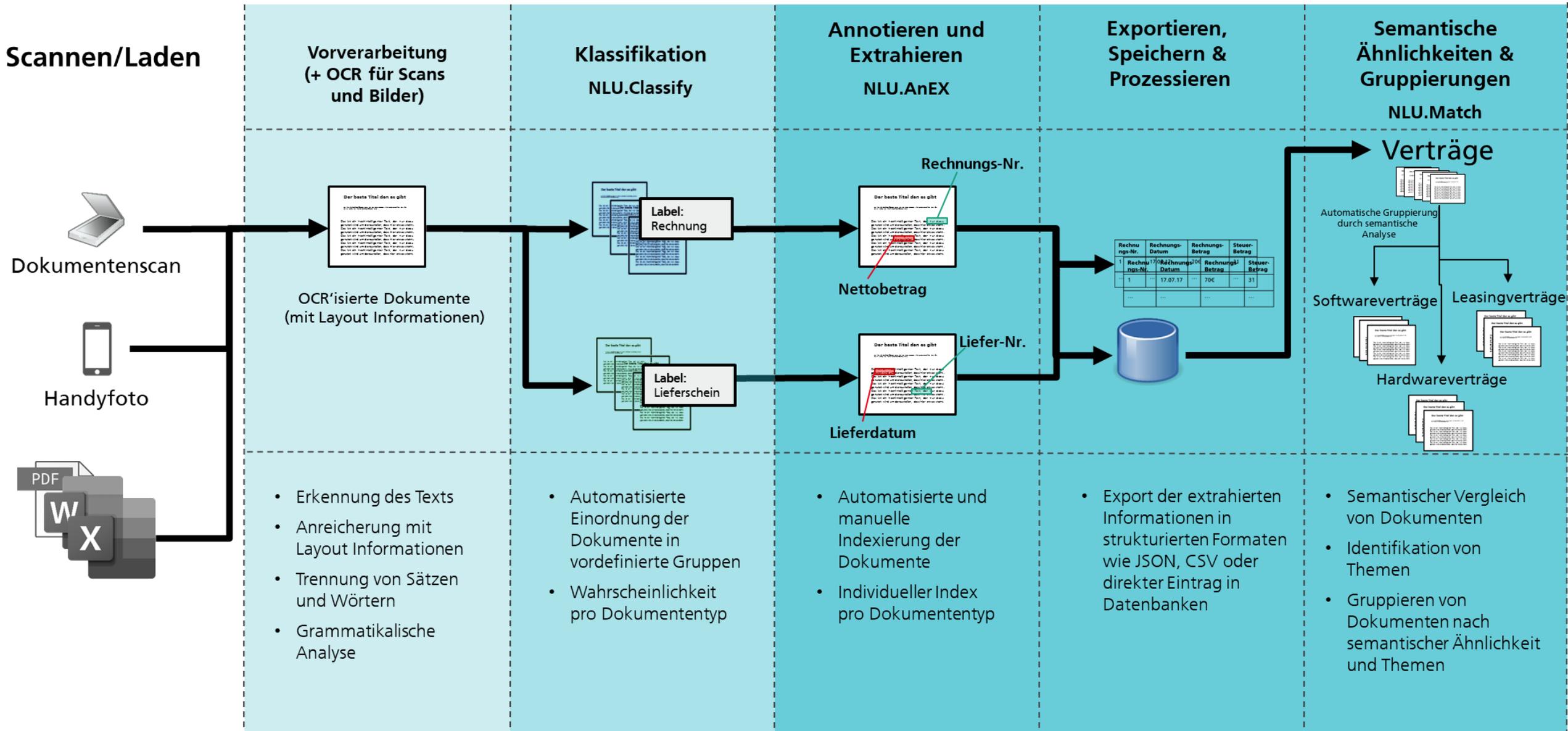
- Wir „fine-tunen“ verfügbare vortrainierte Modelle
 - E.g. BERT, RoBERTa, XLM-R, T5
 - Dies ermöglicht uns rapide, gute Baselines
 - Gelegentlich setzen wir das „Vortrainieren“ auf domänenspezifischen Daten fort
- Wir trainieren eigene Sprachmodelle
 - um domänenspezifisches Vokabular zu berücksichtigen
 - um längere Kontexte zu berücksichtigen
 - weil keine deutschsprachigen oder multilingualen Varianten verfügbar sind
 - Beispiele:
 - Deutschsprachige Longformer oder eigenes deutschsprachiges GPT-Modell

Projektbeispiel: CV-Parsing

Finetuning vorhandener Sprachmodelle

- Workflow aus verschiedenen vortrainierten, multilingualen BERT-Modellen die auf bestimmte Tasks „gefinetuned“ wurden
- Fine-Tuning auf:
 - Named Entity Recognition
 - Erkennen von Entitäten wie Geburtsdaten, Namen, Firmennamen, Jobbezeichnungen, Projektname
 - Paragraphsegmentierung über „Question Answering“
 - „Wie lautet die Beschreibung des Projekts [project_title] bei Firma [company_name] während der Periode [period]?“
 - Klassifikation von Sprachkenntnissen

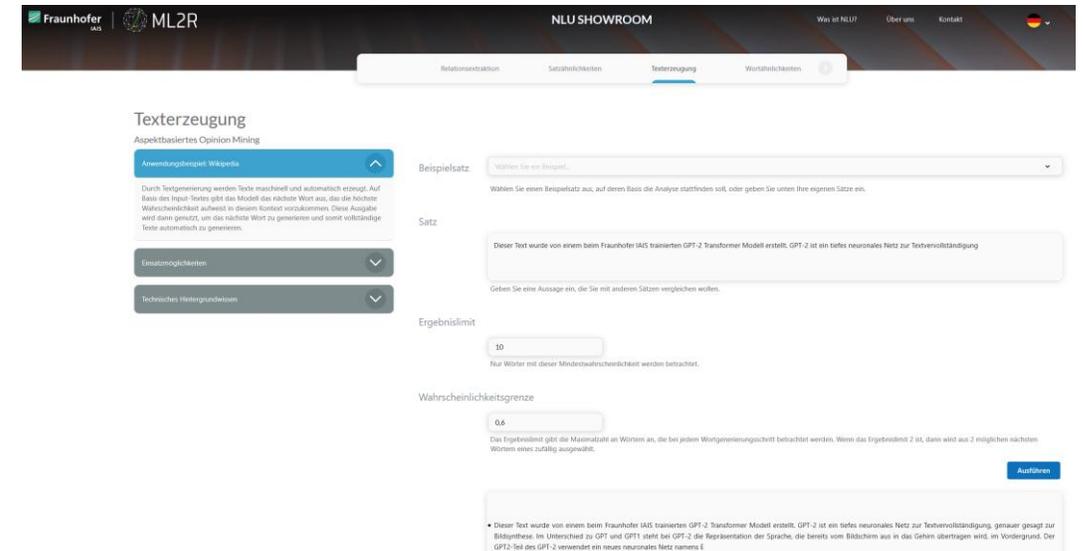




Projektbeispiel: GPT-2

Training neuer Sprachmodelle

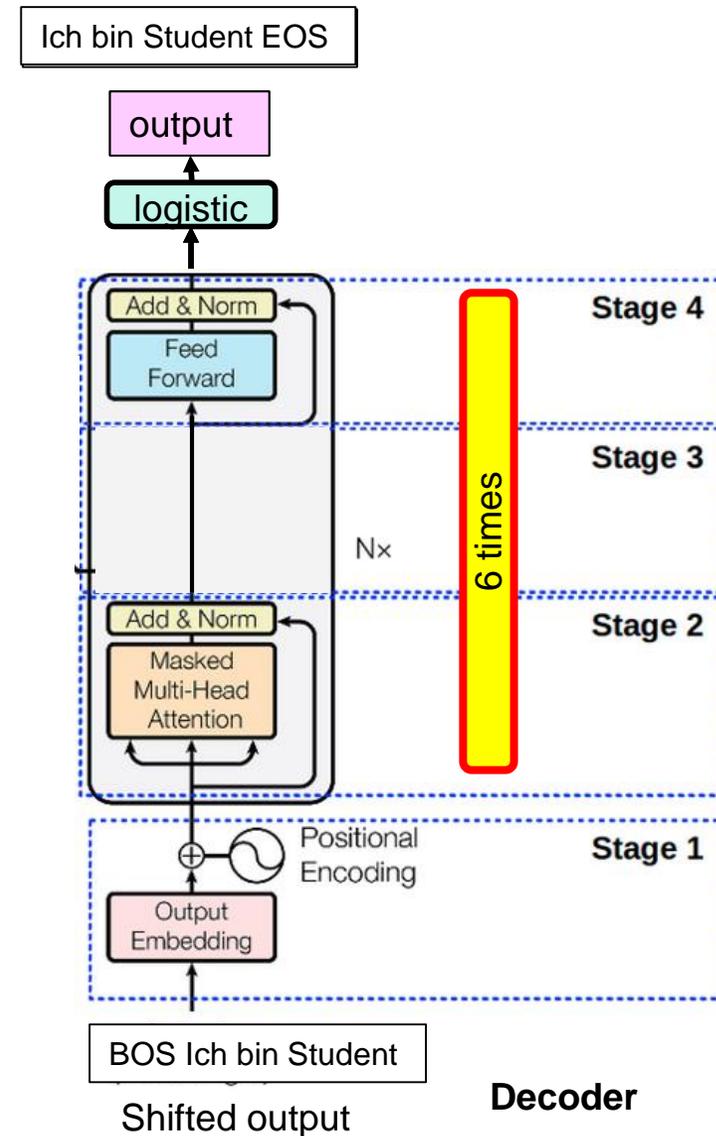
- Training eines „mittelgroßen“ GPT-2 Modells auf deutschsprachigen Texten
- Training auf 8 V100 GPUs für 6 Wochen
- Demo in eigens konzipiertem Showroom



The screenshot displays the 'NLU SHOWROOM' interface. The main navigation bar includes 'Fraunhofer IAIS ML2R' and 'NLU SHOWROOM'. Below the navigation bar, there are tabs for 'Relationsextraktion', 'Satzähnlichkeit', 'Texterzeugung', and 'Wortähnlichkeit'. The 'Texterzeugung' tab is active, showing a section titled 'Texterzeugung' with a sub-heading 'Aspektbasiertes Opinion Mining'. The interface includes a dropdown menu for 'Anwendungsbereich: Wikipedia', a text input field for 'Beispielsatz', a 'Satz' input field, an 'Ergebnislimit' input field set to '10', and a 'Wahrscheinlichkeitsgrenze' input field set to '0.6'. A blue 'Ausführen' button is located at the bottom right. The output area shows a generated sentence: 'Dieser Text wurde von einem beim Fraunhofer IAIS trainierten GPT-2 Transformer Modell erstellt. GPT-2 ist ein tiefes neuronales Netz zur Textervollständigung.' Below this, there is a note: 'Dieser Text wurde von einem beim Fraunhofer IAIS trainierten GPT-2 Transformer Modell erstellt. GPT-2 ist ein tiefes neuronales Netz zur Textervollständigung, genauer gesagt zur Bildsynthese. Im Unterschied zu GPT und GPT1 steht bei GPT-2 die Repräsentation der Sprache, die bereits vom Bildschirm aus in das Gehirn übertragen wird, im Vordergrund. Der GPT2-Teil des GPT-2 verwendet ein neues neuronales Netz namens E'.

GPT2: Transformer Decoder

- Nutzt die Decoder-Schichten des Transformers als Sprachmodell
- Entfernt die Decoder-Encoder attention aus dem Transformer-Modell
- GPT2: 1.5 Mrd. Parameter
BERT: 340 Mio Parameter
- Trainiert auf 40 GB Text
Texte verlinkt von Reddit mit mindestens 3 Karma.



Sprachmodell GPT2

- state-of-the-art Ergebnisse in 7 von 8 Sprachmodellierungs-Tasks.
- **Auf keinem** dieser Datensätze trainiert

Data		SOTA	GPT2
LAMBADA	acc	59.2	63.2
ChildBT-CN	acc	85.7	93.3
ChildBT-NE	acc	82.3	89.1
WikiText2	perplex	39.1	18.3
PTB	perplex	45.5	38.8
enwik8	bpc	0.99	0.93
text8	bpc	1.08	0.98
WikiText103	perplex	18.3	17.5
1BWords	perplex	21.8	42.2

Zero-shot learning

Input

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

Generiert von GPT2

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information.

According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

“The safety of people, the environment and

GPT-3

„Language Models are Few-Shot Learners“

■ Trainiert als Sprachmodell

■ Kein spezifisches Fine-Tuning

■ In-Context Lernen: Lösungsweg im Input und löst das Problem

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush giraffe => girafe peluche ←
5 cheese => ..... ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



g und ggf. Lösungsbeispiele

GPT-3

„Language Models are Few-Shot Learners“

■ Taskübergreifend gilt:

- Je größer das Modell (oft klar besser)
- Beispiele im Input verkettend
 - In-context Lernen
- Andere few-shot Modelle

■ Einige SOTA-“finetuned“ Modelle geschlagen

■ Modell kann sogar einfach in natürlicher Sprache durchführen

Mehr Beispiele dafür was GPT-3 kann:

https://machinelearningknowledge.ai/openai-gpt-3-demos-to-convince-you-that-ai-threat-is-real-or-is-it/#3_OpenAI_GPT-3_can_steal_jobs_from_Software_Engineer

MODEL: GPT-J-6B
Model on Github

Prompt List
Try a classic prompt evaluated on other models

...
a function that yields the depth of a binary tree
...

TOP-P 0.9 Temperature 0.8

Run the model!

Powered by HuggingFace

RESULT

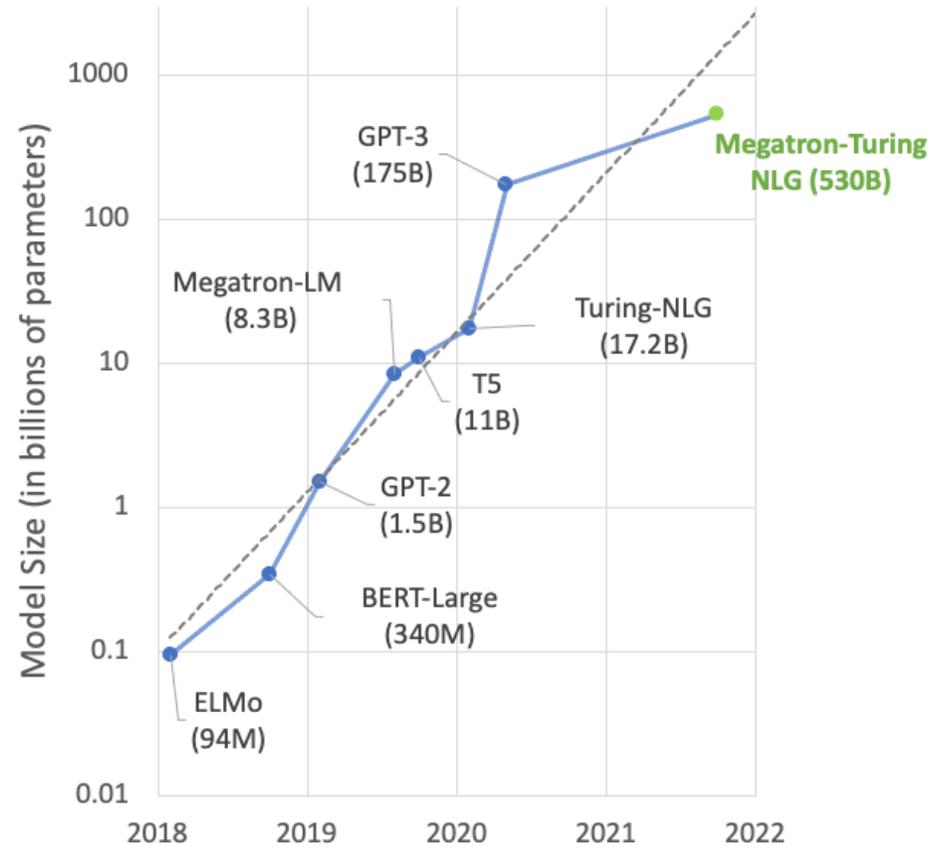
```
...  
a function that yields the depth of a binary tree  
...  
def depth(tree):  
    ...  
    tree: the tree to measure depth  
    returns: the depth of the tree  
    ...  
    if not tree:  
        return 0  
    elif not tree[0]:  
        return 1  
    else:  
        return 1 + max(depth(tree[0]), depth(tree[1]))
```

...ative to Tanzania. An example of a sentence that uses
...w these very cute whatpus.
...and down really fast. An example of a sentence that uses
...my little sister, she got really excited and she
S.
...hat looks like a big pumpkin. An example of a sentence
...ed this yalubalu vegetable that was grown in a garden
...acceleration. An example of a sentence that uses the
...at my father drives to work every day.
...sical instrument. An example of a sentence that uses the
...e me as a gift. I love to play it at home.
...word at it. An example of a sentence that uses the word
screeg is:
We **screeged at each other for several minutes and then we went outside and ate ice cream.**

<https://6b.eleuther.ai/>

Figure 3.16: Representative GPT-3 completions for the few-shot task of using a new word in a sentence. Boldface is GPT-3’s completions, plain text is human prompts. In the first example both the prompt and the completion are provided by a human; this then serves as conditioning for subsequent examples where GPT-3 receives successive additional prompts and provides the completions. Nothing task-specific is provided to GPT-3 other than the conditioning shown here.

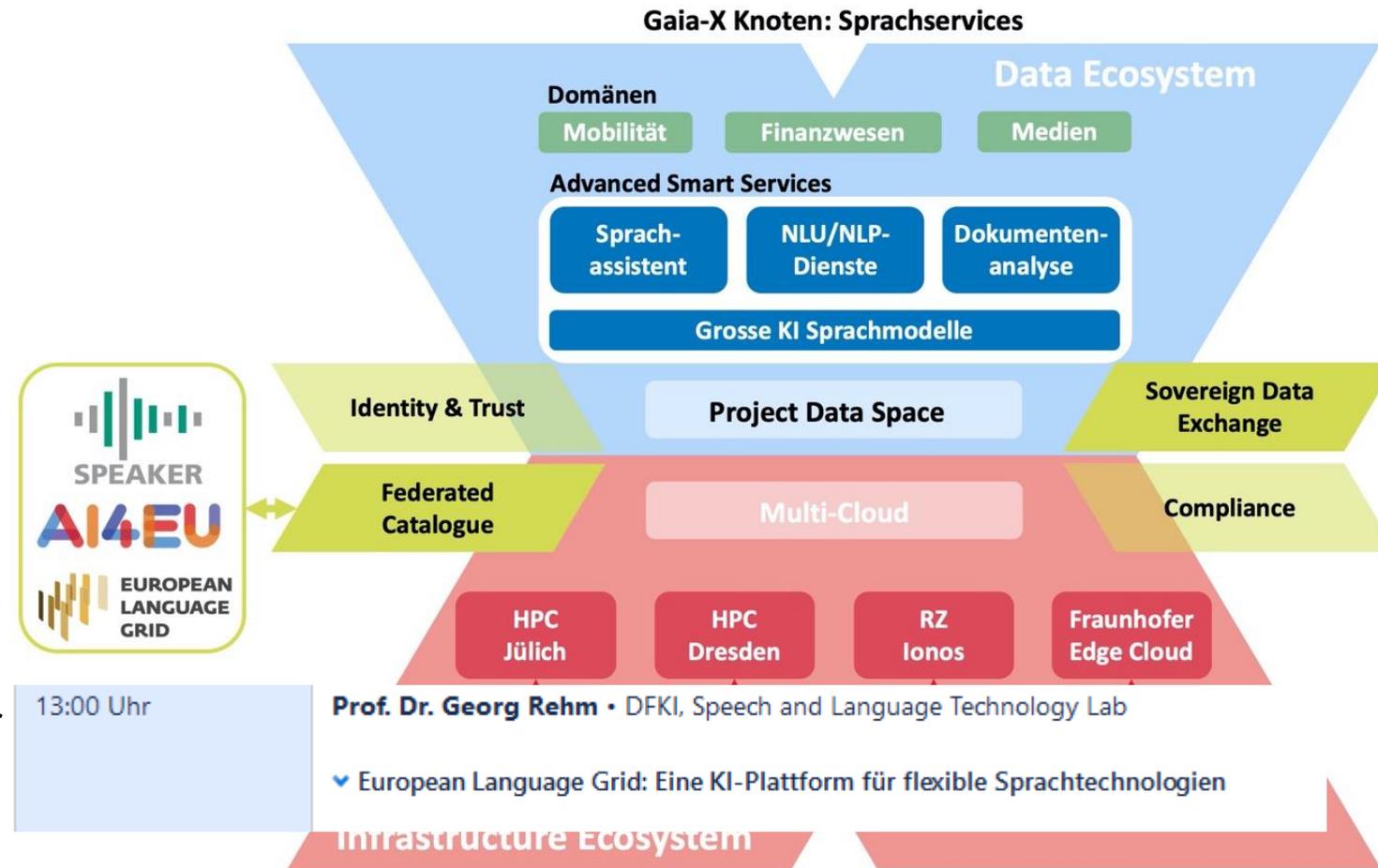
Die Größe vortrainierter Sprachmodelle



[Source: Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model | NVIDIA Developer Blog](#)

OpenGPT-X: Aufbau eines Gaia-X Knotens für große KI-Sprachmodelle und innovative Sprachapplikations-Services

- Erstellung von großen KI-Sprachmodellen basierend auf vertrauenswürdiger Integration der Unternehmensdaten durch Gaia-X
 - Erzeugung **offener Sprachmodelle** und deren Einsatz als **Advanced Smart Services** in ausgewählten **Gaia-X Domänen**
 - **Digitale Souveränität** durch die Erstellung von **KI-Sprachmodellen** => **praxisnahe Anwendung** durch Projektpartner
 - Komplementäres Konsortium: von Infrastruktur über KI-Forschung bis hin zur Anwendung durch Industriepartner
 - **Projektvolumen: 19 Mio. €**



Ein leistungsfähiges Konsortium stellt den Projekterfolg sicher

10 Verbundpartner, 9 assoziierte Partner, 1 Unterauftrag

Das OpenGPT-X Konsortium

- › Koordinations- und KI-Know How: **Fraunhofer IAIS/IIS**
- › Großunternehmen: **IONOS, ControlExpert**
- › KMU, Start-Ups: **aleph alpha, Alexander Thamm GmbH**
- › Rundfunksender: **WDR**
- › Forschungsreinrichtungen: **Fraunhofer, DFKI, Forschungszentrum Jülich, TU-Dresden**
- › Networking: **KI Bundesverband, UnternehmerTUM**
- › Zahlreiche assoziierte Partner, die das Ökosystem stark unterstützen (eco-Verband, Eclipse Foundation, Aalto University, ...)



Aktuelle Forschungsthemen: Reduzieren der Modellgröße

- Transformer-basierte Modelle benötigen viel Arbeitsspeicher und lange Trainingsdauer
- Es bietet sich an die Größe der Modelle zu reduzieren, z.B. durch
 - Komprimieren und/oder reduzieren der Parameter (Shen et al. 2019, Raganato et al. 2020)
 - Durch „Wissensdestillierung“ in kleinere Netzwerke (Sanh et al. 2019, Jiao et al. 2019)
 - Durch Optimierung der Modellgröße über „neural architecture search“ (He et al. 2019)

Table 1: **DistilBERT retains 97% of BERT performance.** Comparison on the dev sets of the GLUE benchmark. ELMo results as reported by the authors. BERT and DistilBERT results are the medians of 5 runs with different seeds.

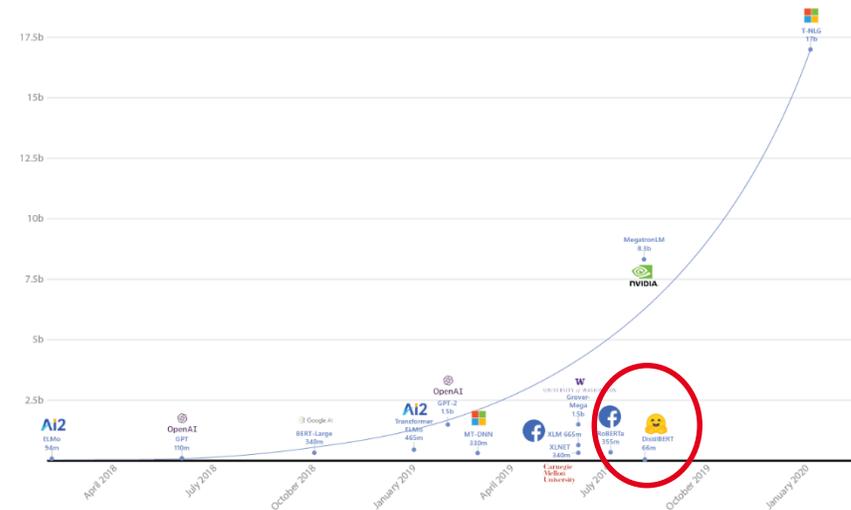
Model	Score	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Table 2: **DistilBERT yields to comparable performance on downstream tasks.** Comparison on downstream tasks: IMDb (test accuracy) and SQuAD 1.1 (EM/F1 on dev set). D: with a second step of distillation during fine-tuning.

Model	IMDb (acc.)	SQuAD (EM/F1)
BERT-base	93.46	81.2/88.5
DistilBERT	92.82	77.7/85.8
DistilBERT (D)	-	79.1/86.9

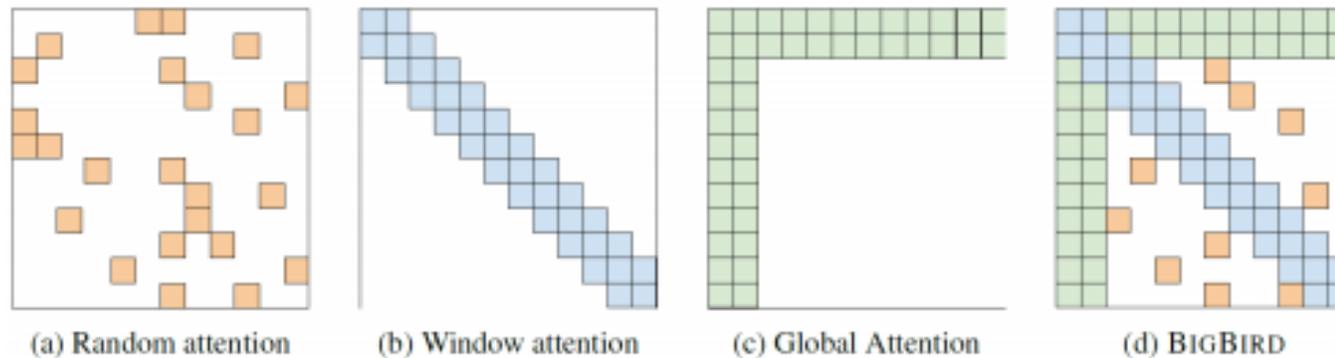
Table 3: **DistilBERT is significantly smaller while being constantly faster.** Inference time of a full pass of GLUE task STS-B (sentiment analysis) on CPU with a batch size of 1.

Model	# param. (Millions)	Inf. time (seconds)
ELMo	180	895
BERT-base	110	668
DistilBERT	66	410



Aktuelle Forschungsthemen: Größere Kontexte erfassen

- Self-attention hat eine hohe Zeit- und Speicherkomplexität, wodurch es schwer ist Transformer auf große Kontexte zu trainieren
- Viele aktuelle Forschungsarbeiten versuchen dies über veränderte Attention-Mechanismen zu verbessern, z.B.
 - Sparse attention die nur bestimmte Inputs berücksichtigt (Beltagy et al. 2020, Zaheer et al. 2020)
 - Gelernte Attention, die ähnliche Token clustered (Kitaev et al. 2020, Roy et al. 2020)
 - Globale Memory Attention welche entfernte Token durch globale Token repräsentiert (Ainslie et al. 2020)



Aktuelle Forschungsthemen: Multimodale Modelle

16:00 Uhr

Hans-Jörg Schäuble · Aleph Alpha GmbH

> Was kommt nach supervised learning? Chancen & Nutzen und ein Blick in die (nahe) Zukunft.

<https://huggingface.co/spaces/anton-l/rudall-e>

Kinda-English ruDALL-E

■ Text steht

ruDALL-E is a 1.3B params text-to-image model by SberAI (links at the bottom). This demo uses an English-Russian translation model to adapt the prompts. Try pressing [Submit] multiple times to generate new images!

■ Aktuelle F

■ Bilder

■ G

■ G

■ Audio

■ E

■ E

■ Video

■ A

Examples

A still life of grapes and a bottle of wine

Город в стиле киберпанк

A colorful photo of a coral reef

A white cat sitting in a cardboard box

What would you like to see?

The German National Library

Clear

Submit

Национальная библиотека Германии

32.48s



log/dall-e/

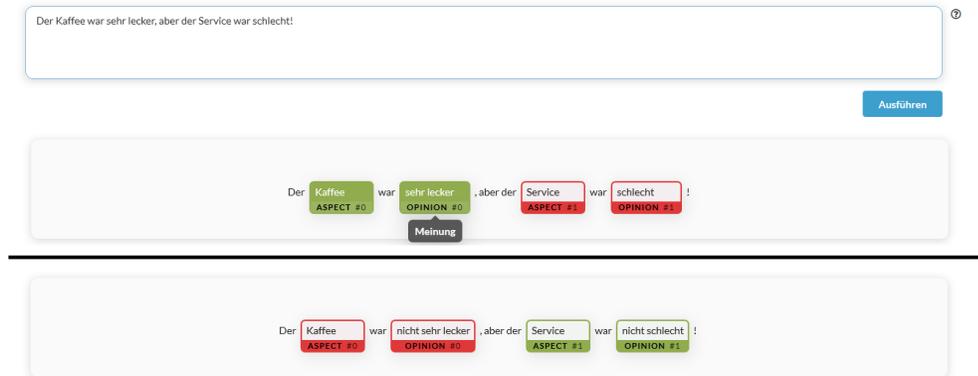
[GitHub](#) | [Article \(in Russian\)](#)

Aktuelle Forschungsthemen: Nutzen zusätzlicher Wissensquellen

- BERT, GPT und Co beziehen ihr Wissen aus statistischen Korrelationen, welche in Texten abgebildet sind
 - Es wird kein, weit-verbreitetes, strukturiertes Wissen genutzt
- Einige Forschungsarbeiten arbeiten an der Integration von „Wissen“ in diesen Modellen:
 - Über Attention auf Wissensgraphen (e.g. Zhang et al. 2019, Peters et al. 2019)
 - Durch Einbindung kontextuell relevanter Texte und Fakten (e.g. Izacard et al. 2020, Karpukhin et al. 2020)
 - Durch hinzufügen logischer Kohärenz (e.g. Clark et al. 2020, Kirsch et al. 2020)

Anwendungsbeispiel: Aspect-Based Opinion Mining – I

- Erkennen von Aspekten und Meinungen einer Bewertung
 - Verknüpfen wenn sie zusammengehören
 - Erkennung der Polarität einer Meinung
- Nutzt KnowBERT und die SenticNet Ontologie (paper pending)
- Zusätzliches „Vortrainieren“ auf deutschsprachigen Restaurantbewertungen
 - Initialisierung von KnowBERT mit dem Bert Modell
- Einbindung zusätzlichen Wissen aus der SenticNet Ontologie über Verknüpfung der Wörter in den Texten
- Vortrainieren von KnowBERT auf den „verknüpften“ Bewertungstexten
- Finetuning des KnowBERT Modells auf Aspekt- und Meinungserkennung, Sentiment Detection, und Aspekt- und Meinungsverknüpfung



Lessons Learned

- Künstliche Intelligenz und Natural Language Understanding werden immer weiter Bestandteil unseres Alltags
- Natural Language Understanding ermöglicht das automatische Verarbeiten und Analysieren textueller Daten
- Unternehmen bieten sich durch NLU viele Möglichkeiten
 - Automatische Klassifikation und Verschlagwortung von Dokumenten
 - Automatische Gruppierung großer Textbestände
 - Automatische Digitalisierung großer Dokumentbestände
 - Automatische Extraktion von Informationen aus Dokumenten
- Das Vortrainieren und „Finetunen“ großer Sprachmodelle ist der Kern vieler NLU-Anwendungen
- Die Forschung befasst sich mit der Anpassung solcher Modelle zur Reduktion der Trainingskomplexität, Erfassung größerer Kontexte, Einbindung multimodaler Inhalte uvm.

„INTELLIGENT SYSTEMS THAT WORK“

Fraunhofer IAIS
Schloss Birlinghoven
53757 Sankt Augustin

www.iais.fraunhofer.de
www.iais.fraunhofer.de/nlu
<https://machinelearning-blog.de/>



Contact

Sven Giesselbach

Team Lead Natural Language Understanding
Fraunhofer IAIS
Telefon +49 2241 14-2249
E-Mail: sven.giesselbach@iais.fraunhofer.de