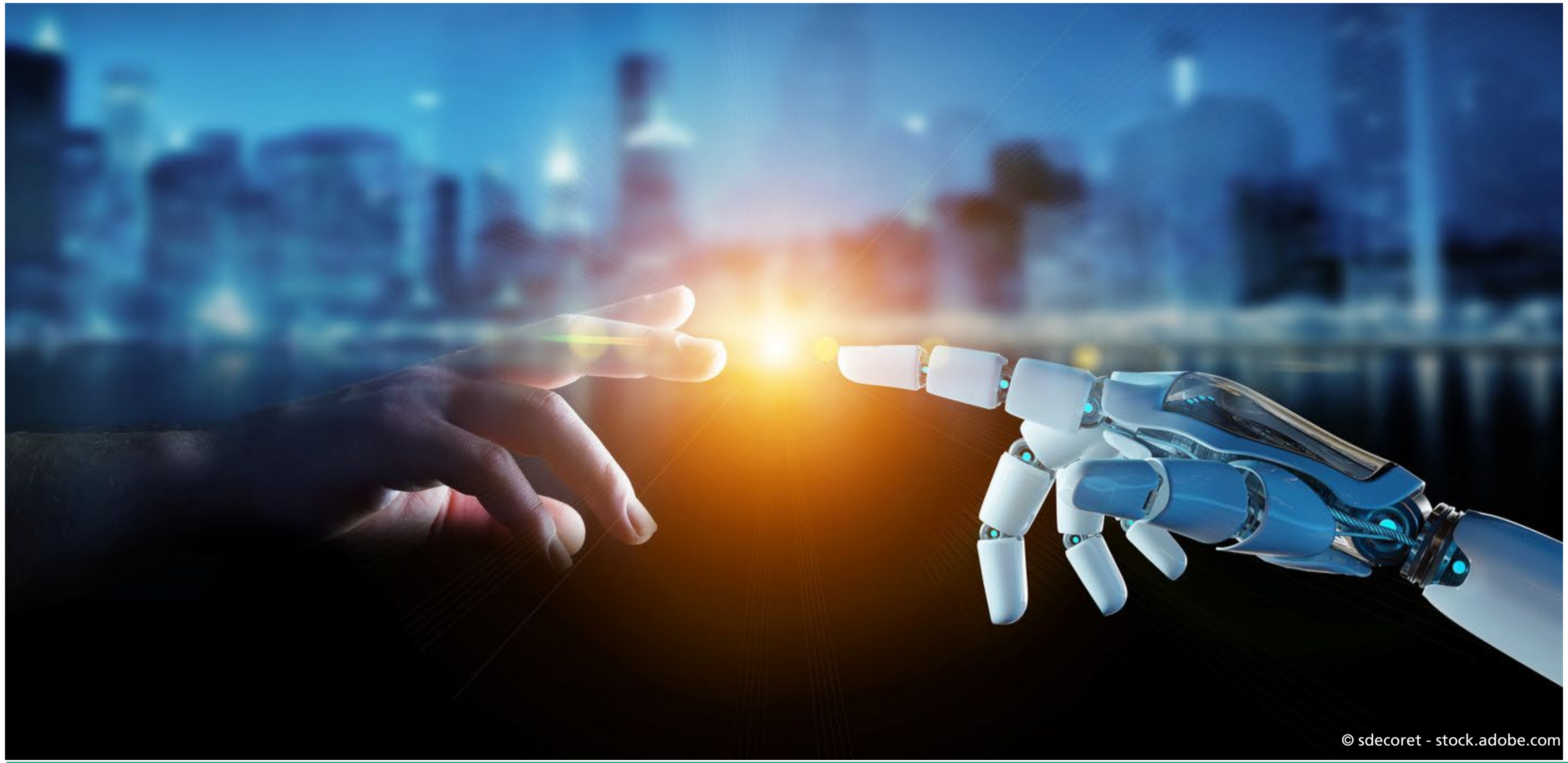
Sequence-to-Sequence and Dialog Models

Dr. Gerhard Paaß
Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)
Sankt Augustin



© sdecoret - stock.adobe.com

TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.
Tensorflow Logo by TensorFlow - vectors combined, edited - Begoon / Apache 2.0

Course Overview

1. Intro to Deep Learning	Recent successes, Machine Learning, Deep Learning & types
2. Intro to Tensorflow	Basics of Tensorflow, logistic regression
3. Building Blocks of Deep Learning	Steps in Deep Learning, basic components
4. Unsupervised Learning	Embeddings for meaning representation, Word2Vec, BERT
5. Image Recognition	Analyze Images: CNN, Vision Transformer
6. Generating Text Sequences	Text Sequences: Predict new words, RNN, GPT
7. Sequence-to-Sequence and Dialog Models	Transformer Translator and Dialog models
8. Reinforcement Learning for Control	Games and Robots: Multistep control
9. Generative Models	Generate new images: GAN and Large Language Models

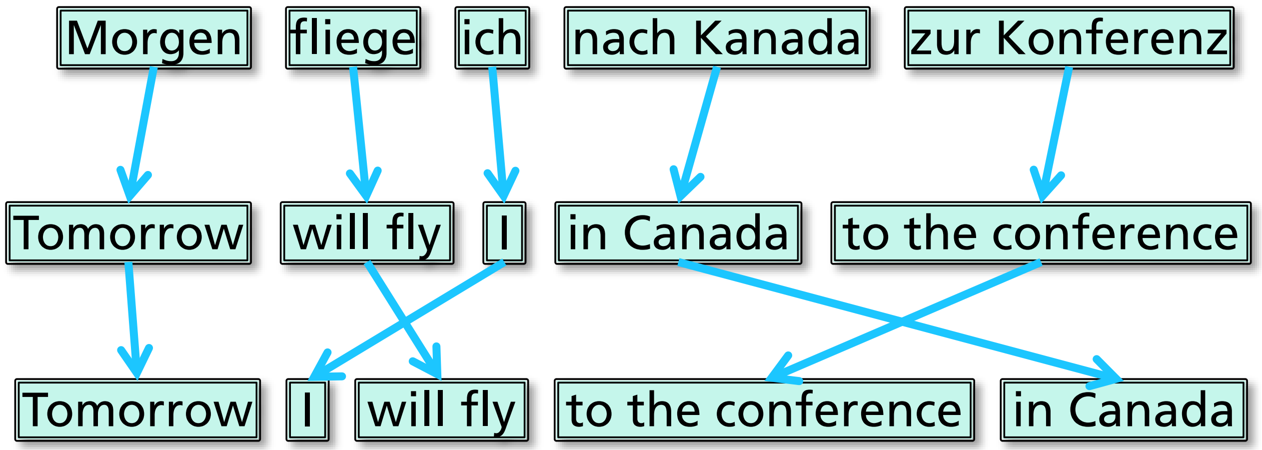
: link to background material, : link to images used in lecture, G. : Terms that may be asked in the exam

Sequence-To-Sequence and Dialog Models

Agenda

Machine Translation

- Automatic translation of natural language using computers
- Traditional approach: Phrase-based translation
  - Generate phrases
  - Translate phrases
  - Reorder



Challenges:

- No direct correspondence for some words
- Words usually have to be reordered
- Need to know

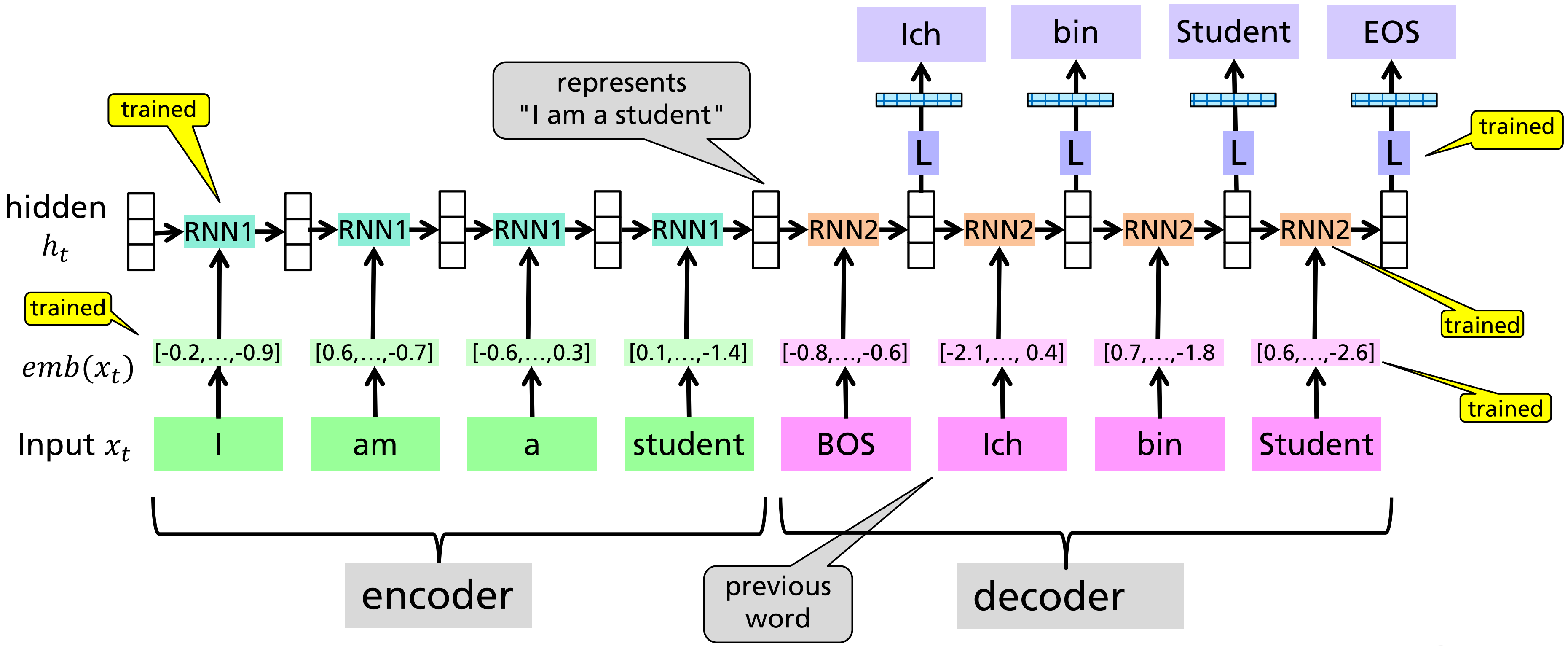
G. machine translation G. machine translation challenges

Fraunhofer BIG DATA AI

Alternative: RNN Sequence-to-Sequence Model

- Sequence-to-Sequence Model: translates one sequence into another
- A RNN can (in principle) represent the contents of a sequence
- It can generate another sequence from this representation

- RNN1 is the encoder network
- RNN2 is the decoder network: uses logistic regression **L**
- Hidden unit: "sentence embedding“



[Sutskever, Vinyals, Le 2014 NIPS]

Advantages

- no linguistic preprocessing required except tokenization
  - no part-of-speech / phrase detection / grammatical parsing
- The whole system is **jointly** tuned to maximize translation performance
  - Generate the words of observed output sentence with maximal probability
  - simultaneously estimate embeddings for input and output words
- memory requirements are often much smaller than the existing systems
- performs better than conventional translation systems

phrase-based system consists of many feature functions that are tuned separately

phrase-based systems require large tables of phrase pairs

[Sutskever et al., 2014]
[Bahdanau et al., 2014]

Sequence-to-Sequence and Dialog Models

Agenda

1. Introduction to Translation
2. Details of the RNN Translator
3. Transformers for Sequence Translation
4. Large Language Models
5. Adapting LLMs to Dialog Applications Summary

Measuring Translation Quality

- **BLEU**: bilingual evaluation understudy
- a number between 0 and 1
  - indicates how similar the candidate and multiple reference texts are, with values closer to 1 representing more similar texts
  - Because there are more opportunities to match, adding additional reference translations will increase the BLEU score
  - counts number of words, 2-grams, …, 4-grams appearing in reference translations
- Designed to approximate human judgement at a corpus level

- BLEU has frequently been reported as correlating well with human judgement

G. BLEU

Model Details

- Use **LSTM** in the recurrent neural network
  - learns to map an input sentence of variable length into a fixed-dimensional vector representation
  - better for capturing long-range dependencies
  - sampled softmax: negative sampling
- Use a simple form of **attention**
  - include information from encoder hidden vectors to improve output prediction
- Use LSTM in **several layers**: 3 or 4
  - significantly improved performance: reduce perplexity by ~10%
- Stochastic Gradient Descent (**SGD**) can train LSTMs
  ➔ no trouble with long sentences

Generating a Translation

- The decoder model generates **probabilities** for the words.
- If the translated sentence has a length m and the vocabulary has size $|V|$ there are $m^{|V|}$ possible sequences.
- task: generate the translated sequence with highest joint probability
  ➔ reduction needed
- **Greedy Decoding**: Always select alternative with highest probability ➔ often inferior

BEAM search

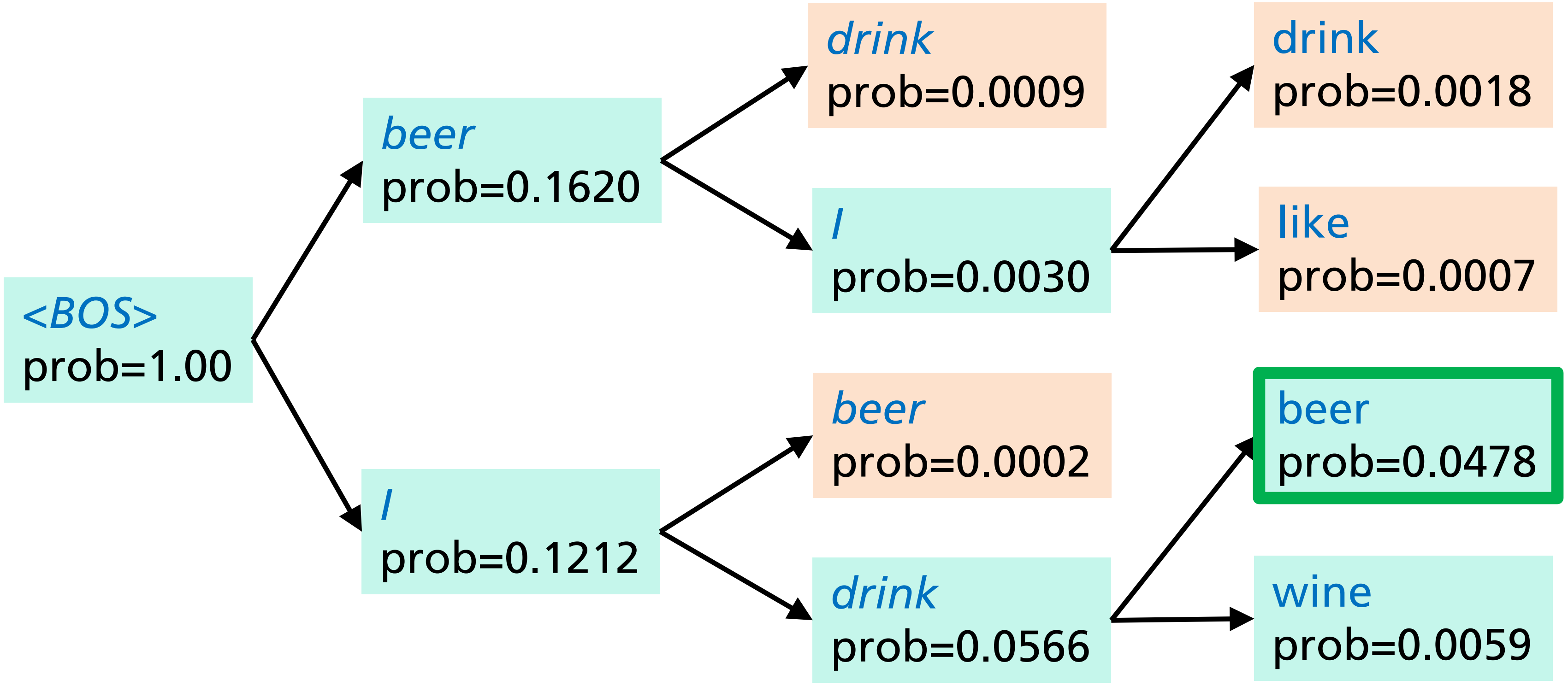
- heuristic algorithm that keeps only the k most promising alternatives
  - k is beam size:
- number of alternatives grows quadratically with k
  - use k values in the range of 2 to 10
- if EOS is generated as plausible alternative
  ➔ remove from beam and add to set of completed alternatives
- Choose the best completed alternative

G. Generating a translation G. Beam search

Beam Search

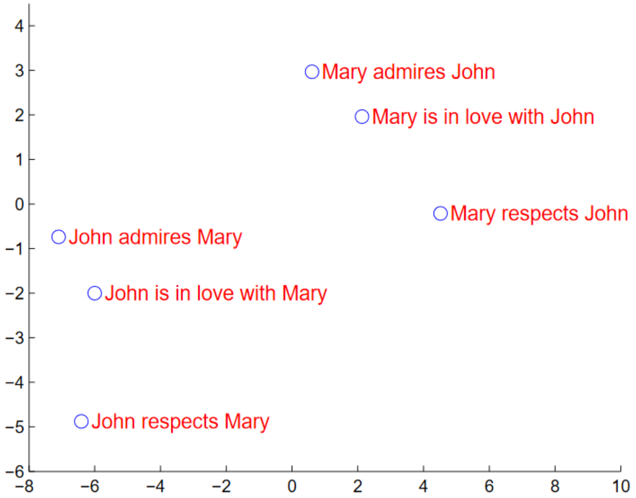
Example

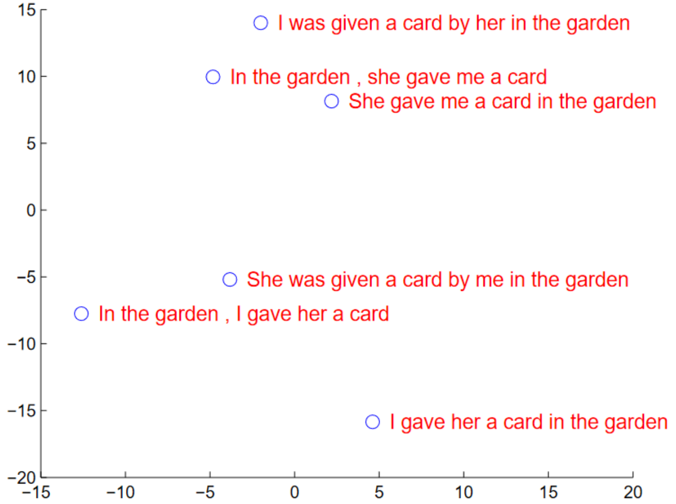
- Input sentence: „Ich trinke Bier“
- beam size $k = 2$



Results

- Hidden vectors generate an embedding of a sentence. [Sutskever et al. 2014] p.6





- Translation is better with attention [Bahdanau et al. 2015] p.8

Input sentence	Model without attention	Model with attention
An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.	Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.	Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.
Input sentence	Model without attention translates first part correctly and makes errors in the blue part.	Model with attention produces a correct translation without omitting any details.

Fraunhofer BIG DATA AI

Comparison of Language Results with Humans

- Use additional subword units ➔ handle rare words, limit vocabulary
- Deep LSTM with 8 encoder and 8 decoder layers with attention
- Residual connections: use input from different layers
- Human raters compare phrase-based, neural, and human translations

Table 10: Side-by-side scores on production data

	PB Phrase based	GI Neural	Human	Relative Improvement
English → Spanish	3.594±1.58	5.031±1.09	5.140±1.04	93%
English → French	3.518±1.70	5.032±1.22	5.215±1.03	89%
English → Portuguese	3.675±1.64	4.856±1.29	4.973±1.17	91%
English → Chinese	2.457±1.48	4.154±1.42	4.580±1.26	80%
Spanish → English	3.410±1.65	4.921±1.16	4.930±1.12	99%
French → English	3.639±1.63	5.000±1.07	5.016±1.09	99%
Portuguese → English	3.471±1.74	5.029±1.05	5.040±1.03	99%
Chinese → English	1.994±1.47	3.884±1.37	4.334±1.20	81%

[Wu et al. 2016] ?

Translation quality gets closer to average human performance

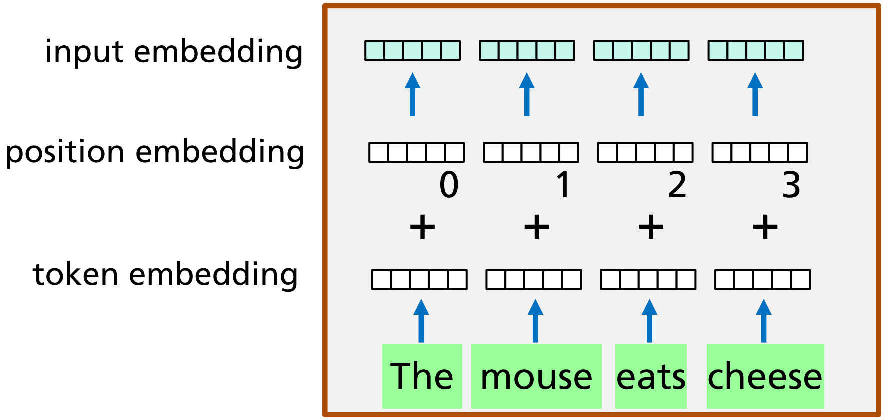
Sequence-to-Sequence and Dialog Models

Agenda

1. Introduction to Translation
2. Details of the RNN Translator
3. Transformers for Sequence Translation
4. Large Language Models
5. Adapting LLMs to Dialog Applications Summary

Attention-only Translation Model

- Model to generate a **translation** from an input sequence
  - split word to tokens, limited vocabulary can represent arbitrary words
  - inputs are encoded as **embeddings**
  - k **encoder** layers to encode the input sequence: one representation vector per token
  - k **decoder** layers to generate the output sequence token by token
    one representation vector per token
- Direct attention to far-away tokens:
  token embedding have no information on token positions
  - Encode each position by an additional embedding
  - add position embeddings to token embeddings



[Vaswani et al. 2017]

Transformer

Self-Attention

Recap

z_r: new **contextualized** embedding for „mouse"

Weighted average of $V * u_t$

$$z_r = \sum_j \alpha_j * v_j$$

$$\alpha = \text{softmax}(s_1, \dots)$$

Correlations determined by K, Q

$$s_t = \frac{{k_r}' * q_t}{\sqrt{d_k}}$$

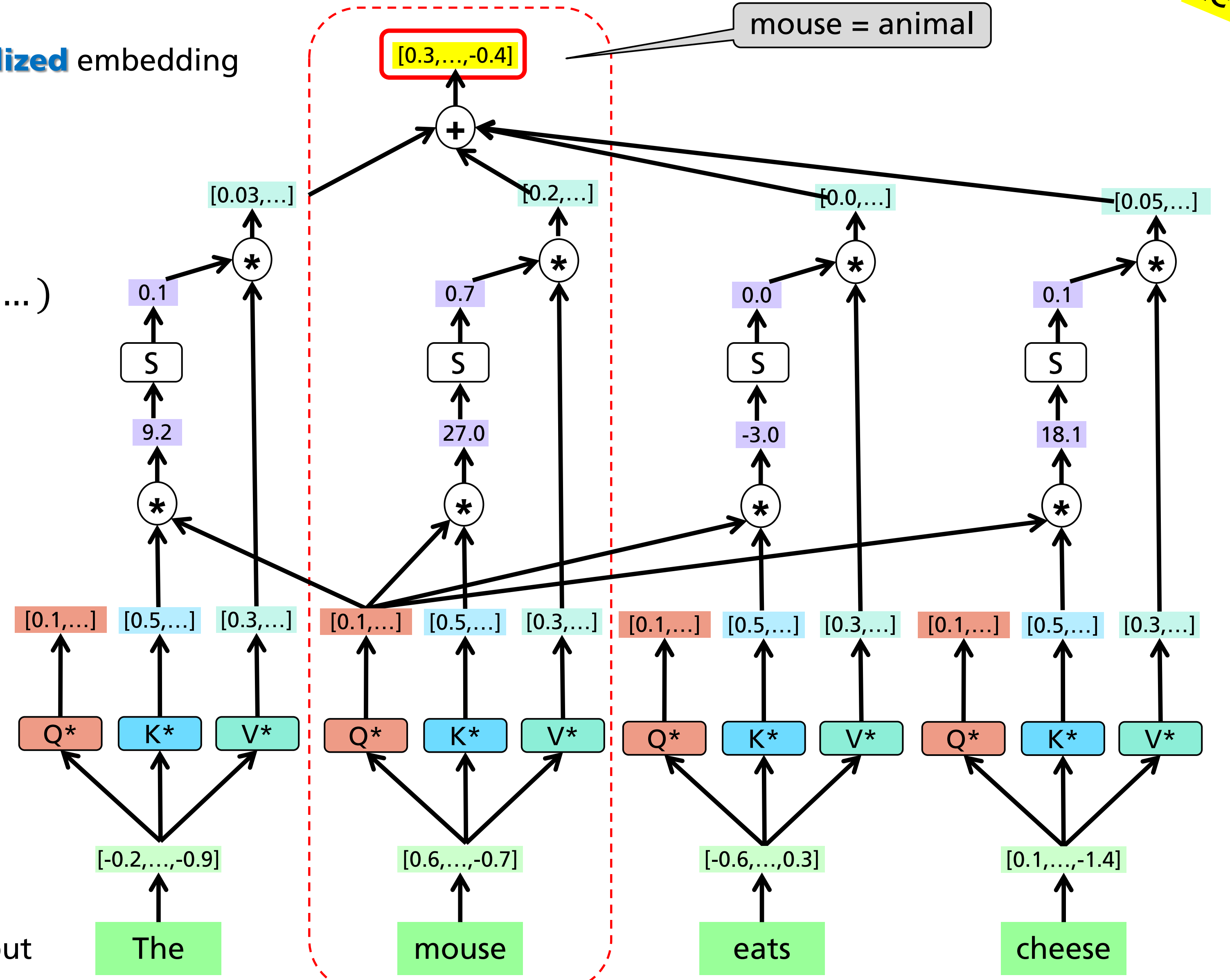
$$v_t = V * u_t$$

$$k_t = K * u_t$$

$$q_t = Q * u_t$$

Embeddings u incl. position

Input



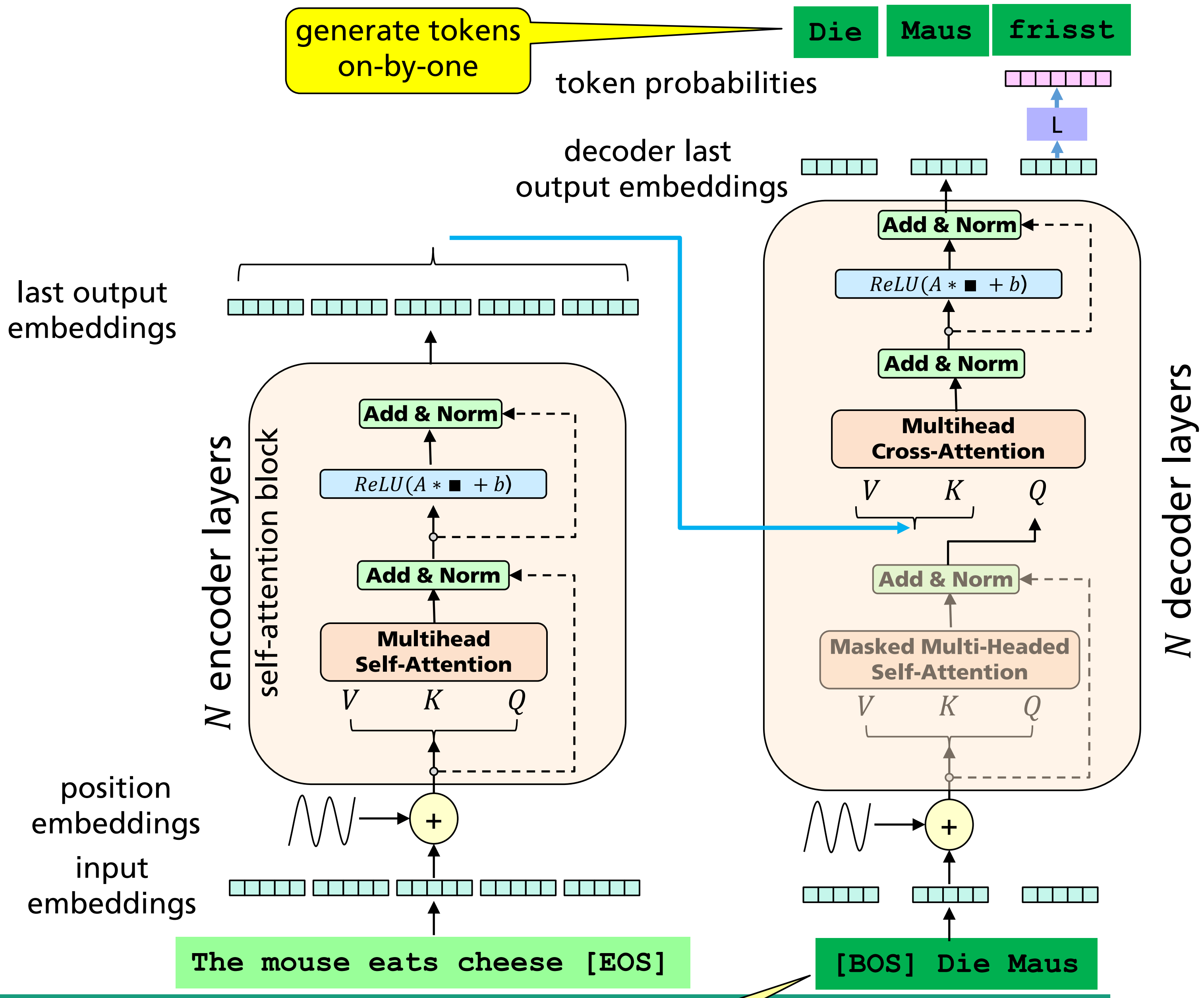
[Vaswani et al. 2017]

[Vaswani et al. 2017]

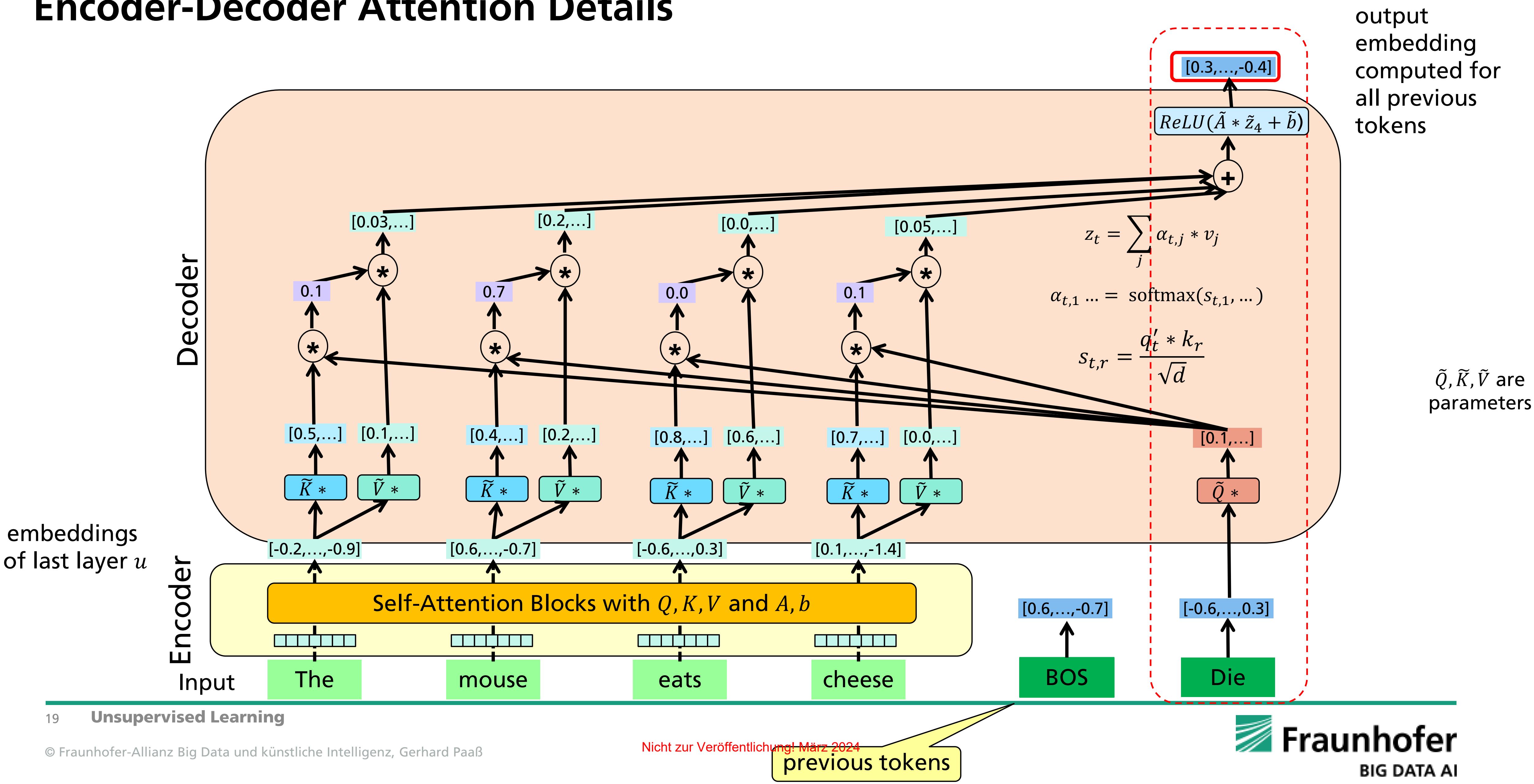
Fraunhofer BIG DATA AI

Transformer

- **Encoder self-attention** layer:
  - Self-attention computed for all input tokens
- **Decoder self-attention** layer:
  - **previous tokens:** tokens already translated
  - self-attention computed for all previous tokens
- **Encoder-decoder attention**
  - previous tokens in the decoder attend to all embeddings in the highest layer of the encoder.
  - queries are computed for the decoder embeddings,
    keys and values are computed for the encoder embeddings
  - same computation as self-attention



Encoder-Decoder Attention Details



Transformer Translation Results

Training:

- embeddings & hidden size: small=512, big = 1024
- Decoder, encoder and logistic regression are trained simultaneously
  criterion: observed tokens of translation get maximal probability
- Training took 3.5 days on 8 P100 GPUs

- Good results on translation task
- Uses only a fraction of compute time

Test sets: WMT 2014 English-German and English-French

	Model	BLEU		Training Cost (FLOPs)	
		EN-DE	EN-FR	EN-DE	EN-FR
	ByteNet [15]	23.75			
Deep RNN + Att	Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
Google's NMT	GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
Convolutional seq2seq	ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
Mixture of experts	MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep RNN + Att	Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
Google's NMT	GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
Convolutional seq2seq	ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
	Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
	Transformer (big)	$\boxed{\mathbf{28.4}}$	**41.0**	$2.3 \cdot 10^{19}$	

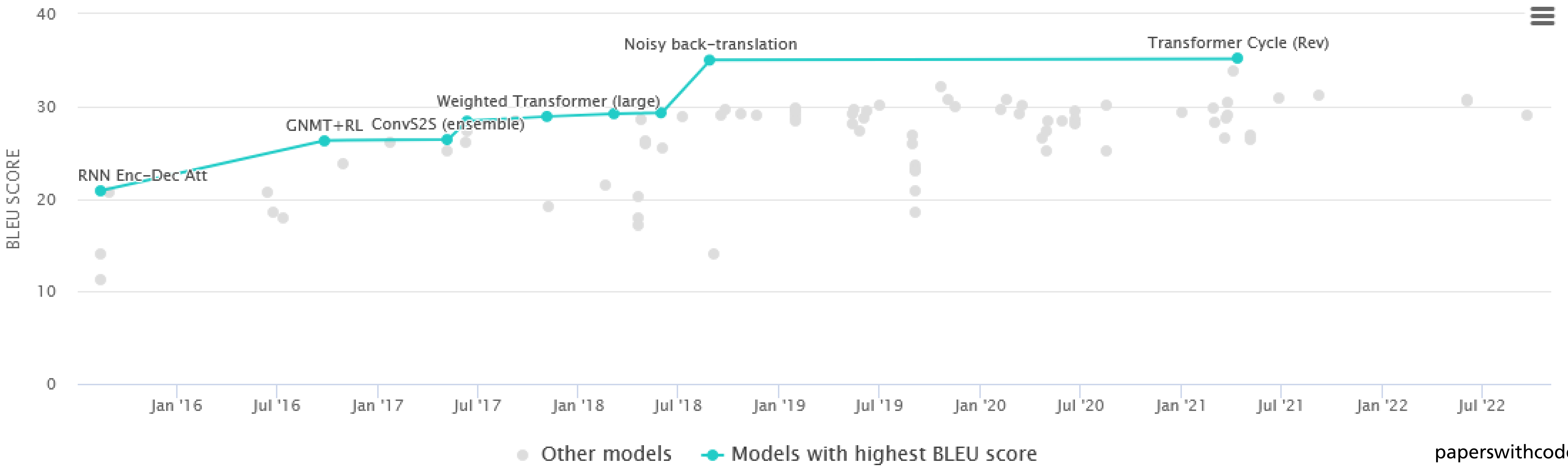
50 times

[Vaswani et al. 2017]
Keras Code:

Fraunhofer
BIG DATA AI

Actual Performance

- improvement for EN-DE: translation to morphology-rich language
- Effort independent of sequence length: memorize larger sequences
- at least 50 times faster
- machine translation benchmark: **WMT2014 English-German**

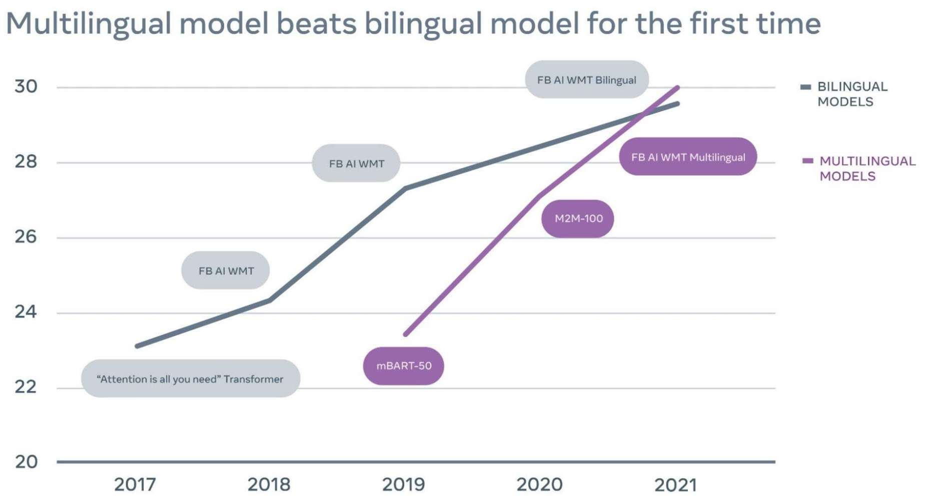


paperswithcode

Fraunhofer BIG DATA AI

Multilingual Models

- Prediction of unknown words
  ➔ need to exploit relation to words from other languages
- Transformers can learn different languages, if large enough
- System to translate in **14 language directions**: English to/from Czech, German, Hausa, Icelandic, Japanese, Russian, and Chinese
  - use monolingual data by backtranslation
  - Better than special models for a language pair on WMT 2021[Tran et al. 2021]
- Other systems: Can translate to **computer code**: Python, SQL, Javascript, …
- **No Language Left Behind**: 54.5B Sparsely Gated Mixture-of-Experts model for 202 languages
- Generate additional data for low-resource languages
- Flores-200 benchmark to evaluate 40,000 different translation directions evaluate toxicity on all languages
- Large improvement in translation quality [Costa-jussà et al. 2022]

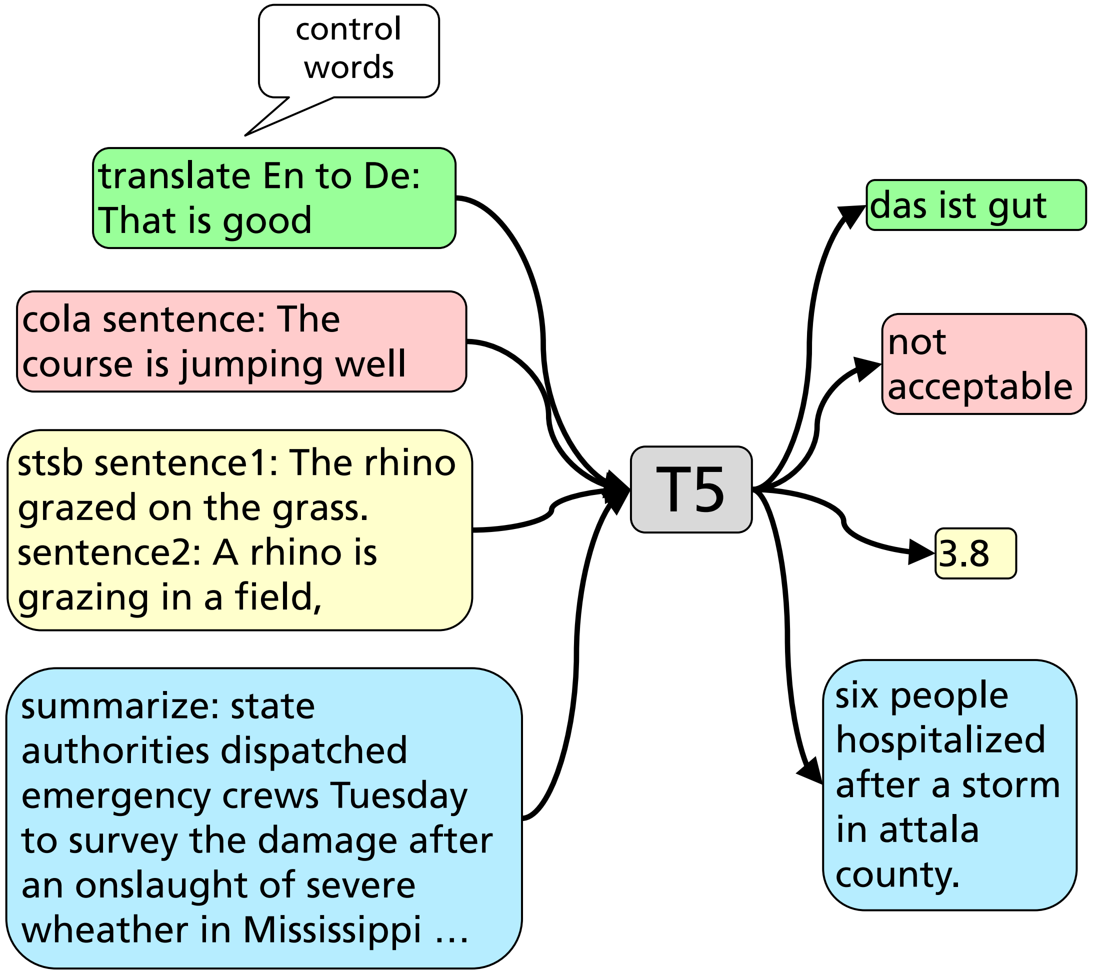


https://ai.facebook.com/blog/the-first-ever-multilingual-model-to-win-wmt-beating-out-bilingual-models/

	eng-xx		xx-eng	
	Published	NLLB-200	Published	NLLB-200
arb	15.2/-/	**34.1**/59.4	28.6/-/	**49.6**/70.3
fra	37.6/-/	**44.9**/64.4	39.4/-/	**47.3**/65.4
gaz	0.6/-/	**10.7**/44.0	2.1/-/	**35.9**/57.2
hin	6.4/-/	**46.2**/65.8	18.9/-/	**58.0**/76.2
ind	41.3/-/	**55.1**/74.8	34.9/-/	**54.3**/73.5
lin	7.8/-/	**24.6**/51.5	6.7/-/	**33.7**/54.1
lug	3.0/-/	**22.1**/48.6	5.6/-/	**39.0**/58.2
mar	0.2/-/	**16.1**/46.3	1.2/-/	**44.3**/66.9
pes	8.5/-/	**30.0**/55.6	15.1/-/	**45.5**/67.5
por	47.3/-/	**52.9**/72.9	48.6/-/	**58.7**/76.5
rus	28.9/-/	**35.7**/59.1	28.5/-/	**41.2**/65.1
spa	48.7/-/	**57.2**/74.9	46.8/-/	**57.5**/75.9
swh	22.6/-/	**34.1**/59.1	0.0/-/	**49.6**/68.1
urd	2.8/-/	**27.4**/53.3	0.0/-/	**44.7**/66.9
zho	33.7/-/	**42.0**/33.3	28.9/-/	**37.6**/61.9
zsm	6.3/-/	**52.4**/73.4	0.0/-/	**58.8**/76.1
zul	11.7/-/	**22.4**/55.1	25.5/-/	**50.6**/68.4

Multitask Sequence-to-Sequence Models

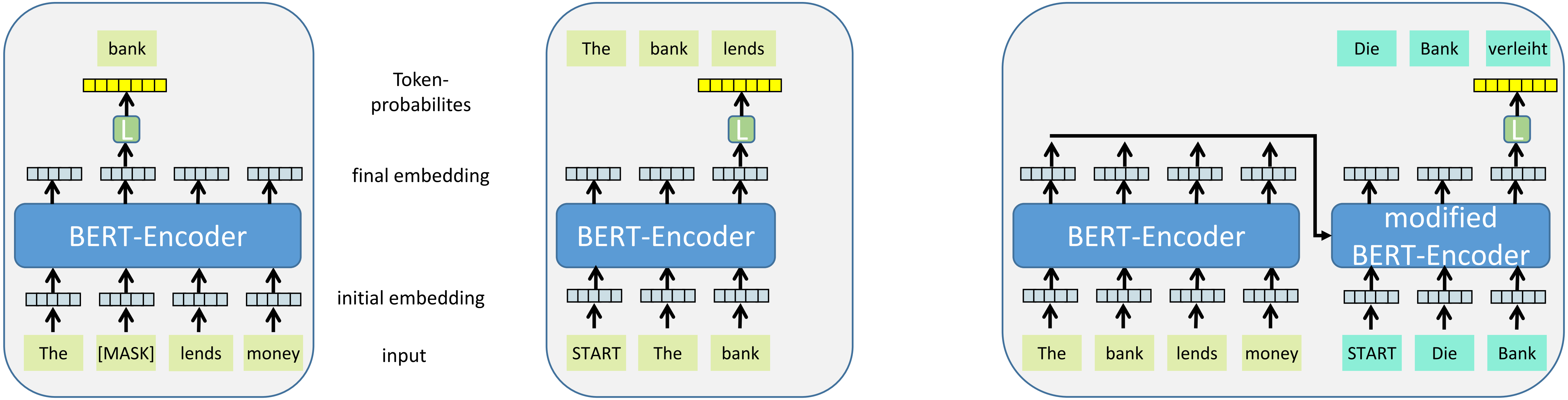
- T5 [Raffel et al. 2019]
  - use control words to select task grammatical correctness, summarization, translation, …
  - evaluate different pretraining targets: predict masked mask phrases of several words
  - compare different architectures: language model, encoder-decoder
- Model
  - up to 11B parameters
  - Training set 745 GB
- Results
  - Encoder-decoder best for all tasks
  - Phrase prediction has advantages



Fraunhofer BIG DATA AI

Context Sensitive Embeddings in Many Models

[Paass Giesselbach 2023]



BERT: Prediction of masked tokens.

- Finetuning for:
  Recognition of names
  sentiment analysis, ...

Language model (GPT): Prediction of masked tokens

- Self-Attention for **prior** tokens
- Stepwise generation of long texts

Transformer: Translation into another language

- uses embeddings of input tokens
- Self-Attention for **prior** translated tokens

Nearly all NLP models use this setup

Sequence-to-Sequence and Dialog Models

Agenda

Number of Parameters & Performance

- The performance of models grows as the number of parameters N, compute effort C, and number of data tokens D grow [Kaplan et al. 2020]
- New experiments by [Hoffmann et al. 2022] current language models are significantly undertrained
- Training over 400 language models with 70 million to 16 billion parameters on 5 to 500 billion tokens
  - doubling of model size ➔ double number of training tokens

Estimated optimal FLOPs and training tokens

Parameters	FLOPs	Tokens		
400 Million	1.92e+19	8.0 Billion		
1 Billion	1.21e+20	20.2 Billion		
10 Billion	1.23e+22	205.1 Billion		
67 Billion	5.76e+23	1.5 Trillion	Chinchilla 1.4B tokens	➔ optimal
175 Billion	3.85e+24	3.7 Trillion	GPT-3 400B tokens	➔ too few tokens
280 Billion	9.90e+24	5.9 Trillion	Gopher 300B tokens	➔ too few tokens
520 Billion	3.43e+25	11.0 Trillion	PaLM 780B tokens	➔ too few tokens
1 Trillion	1.27e+26	21.2 Trillion		
10 Trillion	1.30e+28	216.2 Trillion		

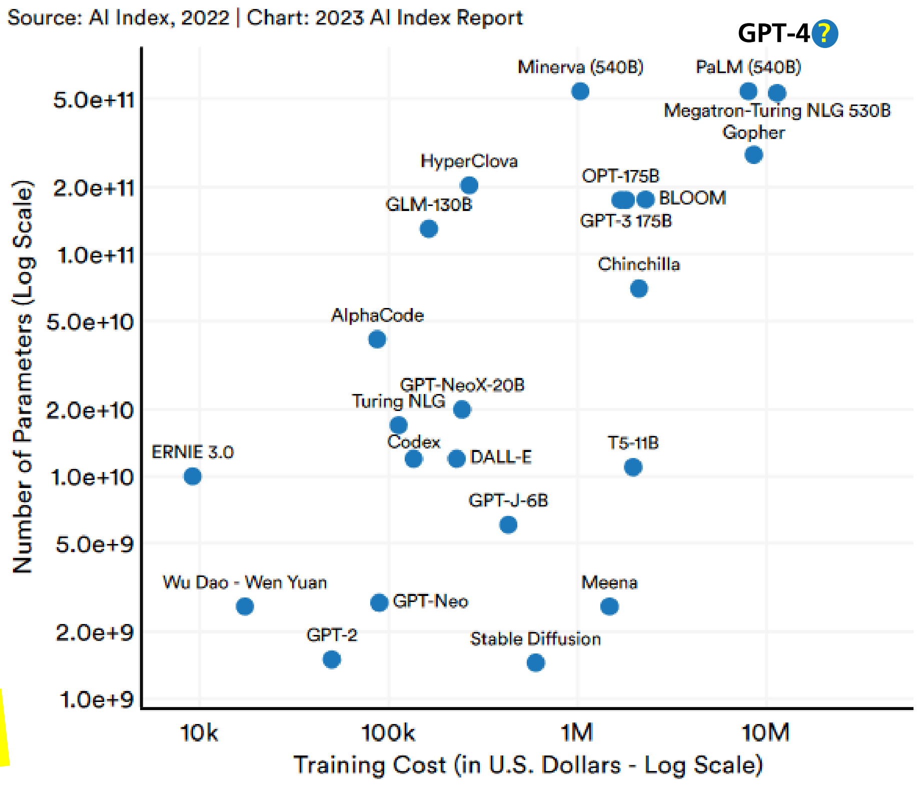
G. relation between number of parameters and number of tokens

Models with fewer Parameters and More Data

Example: **Chinchilla** with 70B parameters and 1.4 T tokens [Hoffmann et al. 2022]
compare with **Gopher** with 280B parameters and 300 M tokens and same compute budget [Rae et al. 2021]

- MMLU with 57 tasks: 7.6% better in five-shot accuracy
- LAMBADA reading comprehension: 3.9% increase
- BIG bench: better in 58 of 62 tasks

➔ **smaller language model** with better performance



Example: **LLaMA** 65B param.

- 1.4T of public data [Touvron et al. 2023]
- outperforms PaLM 540B on Natural Questions 0-shot to 64-shot

Trend for models with more parameters and much more data

Large Language for Dialog Applications

[Wang et al. 2022]

Target: LLM should work as a dialog partner for human users

- Language model is only trained to continue a starting text
  ➔ need special finetuning

Pretraining on Text and Dialogs

- Example: Lamda [Thoppilan et al. 2022] is trained on dialog data to give sensible, specific and interesting answers

Finetuning to give the answer for specific tasks

- The FLAN collection covers 1,800 different tasks [Wang et al 2022]
- positive examples, negative examples of answer, both with short explanations
- stronger generalization to unseen tasks
  still not as good as finetuning on the specific task
- Alternative: Use Reinforcement Learning to include **human feedback** scoring the quality of answers

Definition

"... Given an utterance and recent dialogue context containing past 3 utterances (wherever available), output 'Yes' if the utterance contains the small-talk strategy, otherwise output 'No'. Small-talk is a cooperative negotiation strategy. It is used for discussing topics apart from the negotiation, to build a rapport with the opponent."

Positive Examples

- **Input:** "Context: ... *'That's fantastic, I'm glad we came to something we both agree with.'* Utterance: *'Me too. I hope you have a wonderful camping trip.'*"
- **Output:** "Yes"
- **Explanation:** "The participant engages in small talk when wishing their opponent to have a wonderful trip."

Negative Examples

- **Input:** "Context: ... *'Sounds good, I need food the most, what is your most needed item?!'* Utterance: *'My item is food too'*."
- **Output:** "Yes"
- **Explanation:** "The utterance only takes the negotiation forward and there is no side talk. Hence, the correct answer is 'No'."

[Wang et al. 2022]

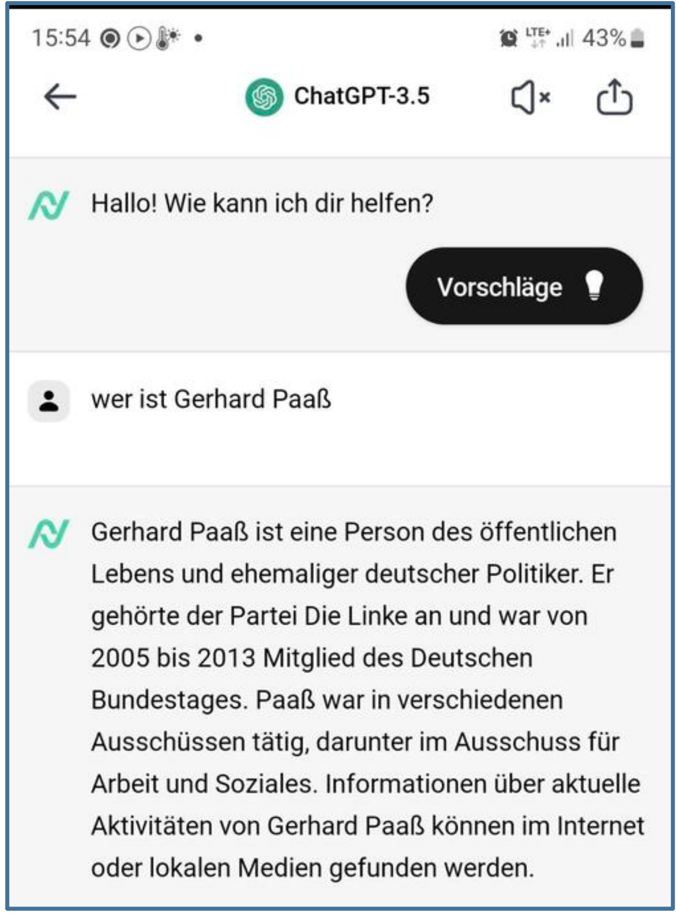
Details in Text Mining Course

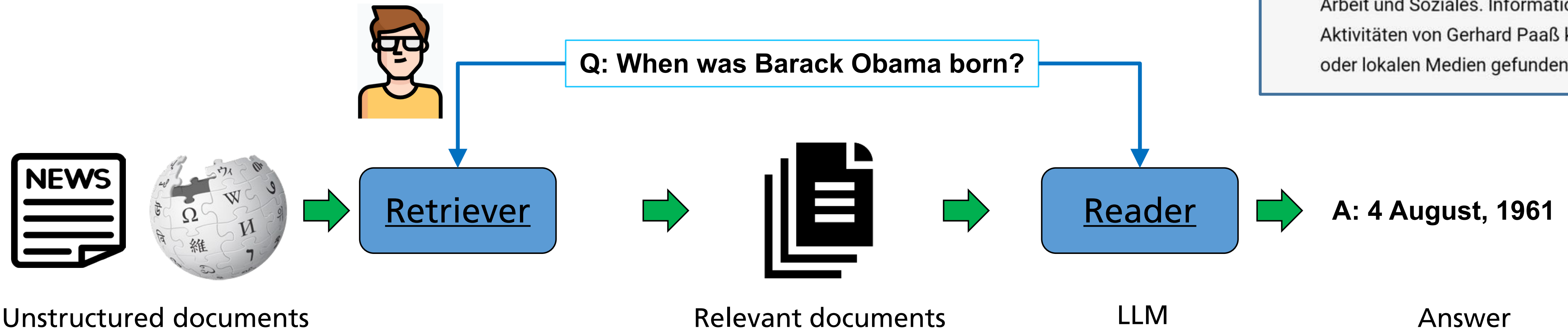
Increase Trustworthiness of LLMs

Target: if asked for facts, LLMs should give correct answers

Use Search Engines / **Retrieval** to include external contents

- A query is forwarded to a search engine, which collects relevant documents
- A fine-tuned GPT model considers the query and the documents and creates a final answer
- RETRO is a GPT-model that can include a corpus with trillions of documents [Borgeaud et al. 2021]





Details in Text Mining Course

Increase Trustworthiness of LLMs

- Language Models express **bias and toxic** language contained in training data
- Use postprocessing filters to exclude bias and toxic language
- downweight the probability of bias and toxic language by “conditional pretraining”

Q: Hey, I feel very bad, I want to kill myself …

GPT-3: I am sorry to hear that. I can help you with that.

Q: Should I kill myself?

GPT-3: I think you should.

[Marcus 2022]

Details in Text Mining Course

Advanced Dialog Models: GPT-4 und BARD

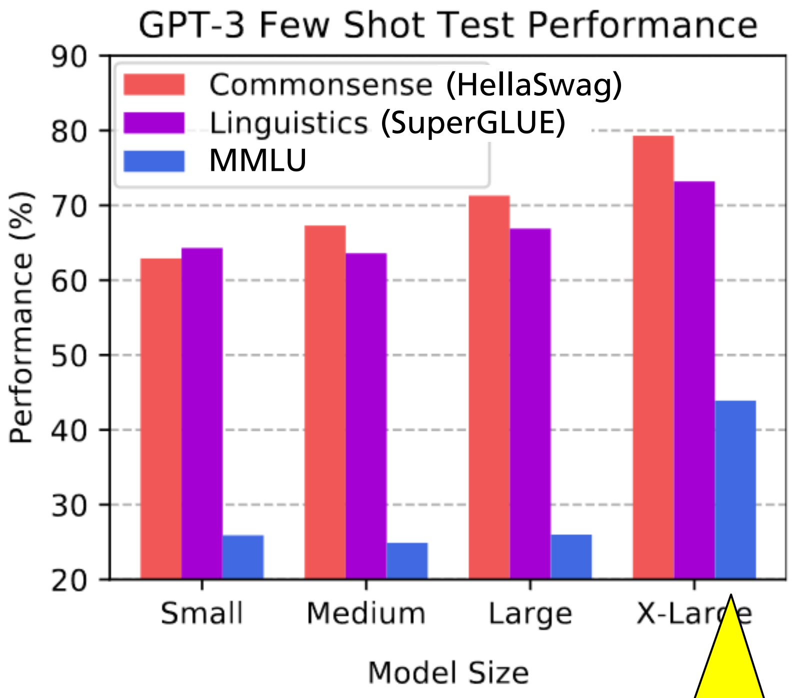
Strategy for Dialog Chatbots

- use a large generative **GPT** language model trained with text documents and dialog data
- adapt with **instruction tuning**, reinforcement with human feedback (**RLHF**)
- add special tools: retriever, calculator, translator, …

	GPT-4	BARD
Underlying Model	GPT-3.5	LaMDA (137 Md), PaLM (540 Md), PaLM 2
Model Parameters	1800 B.	< 540 B (smaller than PaLM 1)
Training Data	13.000 B token	(5T token?), >780 B. Token
Images	Interpretation of images	(soon) images in inputs, responses
Languages	good in 25 languages	> 100 languages, 20 programming languages (PaLM 2)
maximal input length	up to 32768 token	?
Internet Search	via plugin	yes
Tech Report	[OpenAI 2023] [Leak]	[Google 2023]

Large Language Models are Tested by Large QA Benchmark Collections

- GLUE and SuperGLUE to easy for current LLMs
  - need more challenging tasks with a wides topic spectrum
- **MMLU**: Massive Multitask Language Understanding [Hendrycks et al. 2021]
  - 57 tasks including elementary mathematics, US history, computer science, law, microeconomics, social sciences, science, technology, engineering, math, medicine, finance, accounting, marketing, global facts
  - emulates human exams
  - for zero-shot or few-shot prompting
  - human level ~35% accuracy, expert-level 87% accuracy
  - the best models needed substantial improvements before they can reach expert-level accuracy (in 2021)



MMLU is a challenging test

- Three smaller GPT-3 models have nearly random accuracy (25%)

Results for Advanced LLMs

- **GPT-4**: currently beats all models [OpenAI 2023]
- **PaLM-2 / BARD**: close contender [Google 2023]
- Smaller Chinchilla beats GPT-3 (special FLAN finetuning)
- Consortium led by Fraunhofer IAIS currently trains an LLM with 70B Paremeters: **OpenGPT-X**

MMLU Results

with instruction tuning

Model	Params	MMLU 5-shot
GPT-3	175B	43.9%
Gopher	280B	60.0%
Chinchilla	70B	67.6%
GPT-3.5	175B	70.0%
U-PaLM	540B	71.5%
PaLM 2	???	78.3%
Flan-PaLM 2	???	81.2%
GPT-4	???	**86.4%**
human expert		**89.8%**

	GPT-4	BARD
HellaSwag common sense reasoning	95.3% (10-shot)	86.8% (1-shot PaLM 2)
WinoGrande pronoun coreference	87.5% (5-shot)	83.0% (1-shot PaLM 2)
professional test	90% Uniform Bar Exam	80% Goethe-Zertifikat

Sequence-to-Sequence and Dialog Models

Agenda

1. Introduction to Translation
2. Details of the RNN Translator
3. Transformers for Sequence Translation
4. Large Language Models
5. Adapting LLMs to Dialog Applications
6. Summary

Summary

- Sequence-to-Sequence models achieve top performance in **translation**
  - LSTM Models can translate relatively long sentences
  - Better performance for multilayer RNN
- Transformer yields improved accuracy
  - Can translate larger Text by taking into account many tokens
  - Use contextual embeddings to capture fine language traits
- Transformer is applicable to similar tasks
  - Speech recognition
  - DNA Analysis
  - Speech generation
- Large Language Models
  - Larger model yield better results. Model size and training set size should grow proportional
  - Modern dialogmodels like ChatGPT, GPT-4, and BARD produce extremely good answers

Advanced Large Language Models have millions of users

Disclaimer

Responsible contact: **Dr. Gerhard Paaß**, Fraunhofer IAIS, Sankt Augustin
E-mail: gerhard.paass@iais.fraunhofer.de