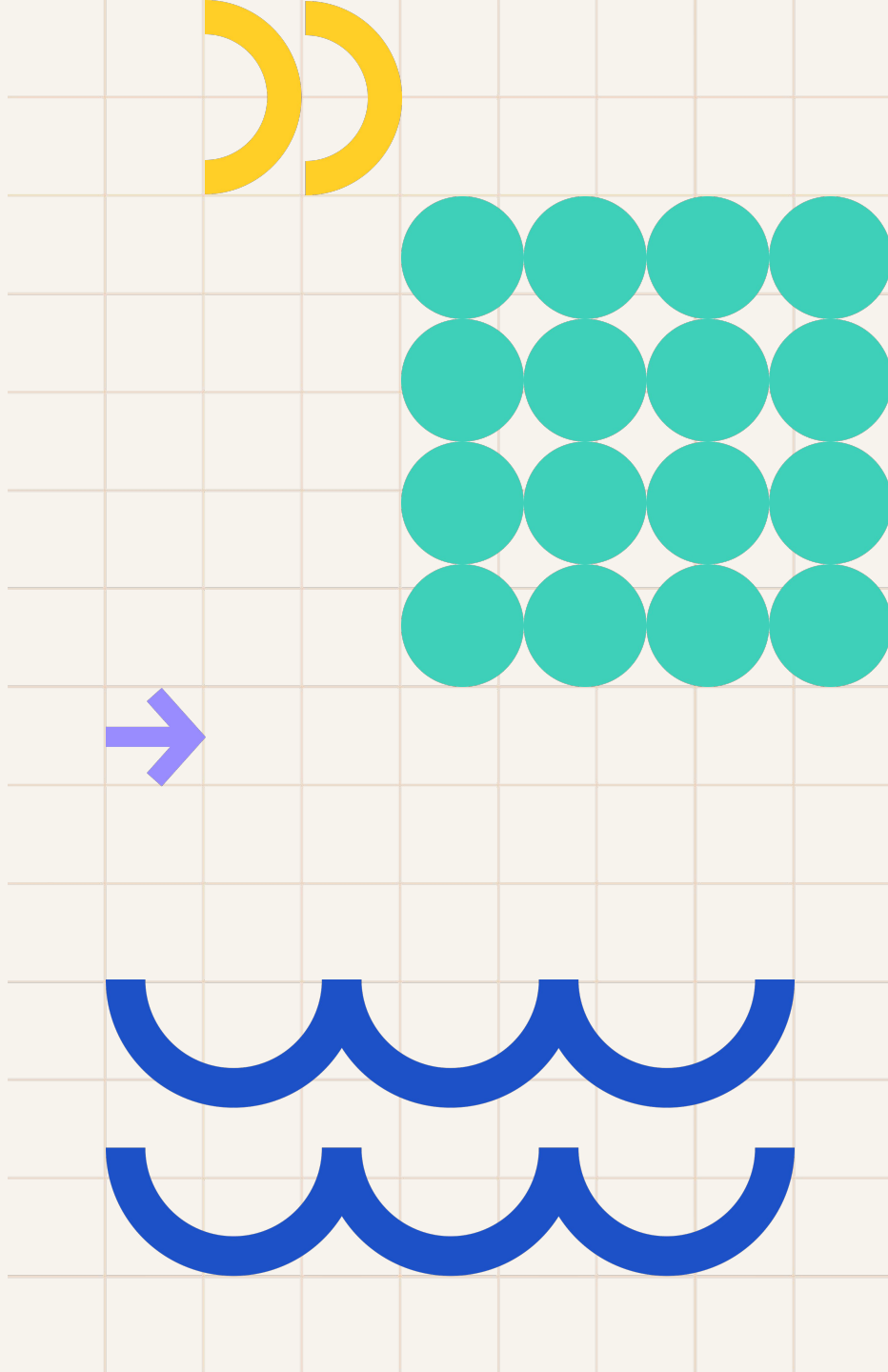
AIRI / Университет Сириус

# Machine Translation and Generative Models

Alsu Sagirova



# Agenda

01 Machine Translation

02 Generative Models

Telegram

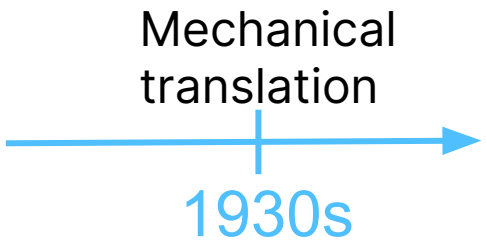
NLP course github

Feedback

# 01

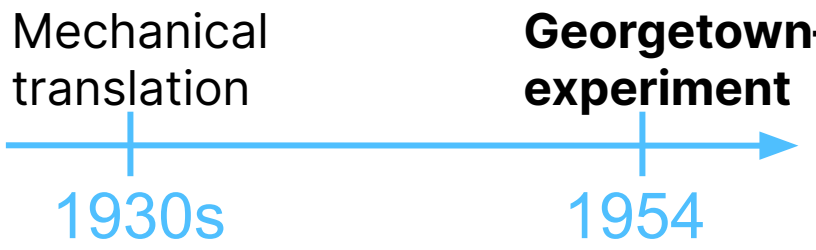
# Machine Translation

# Machine Translation

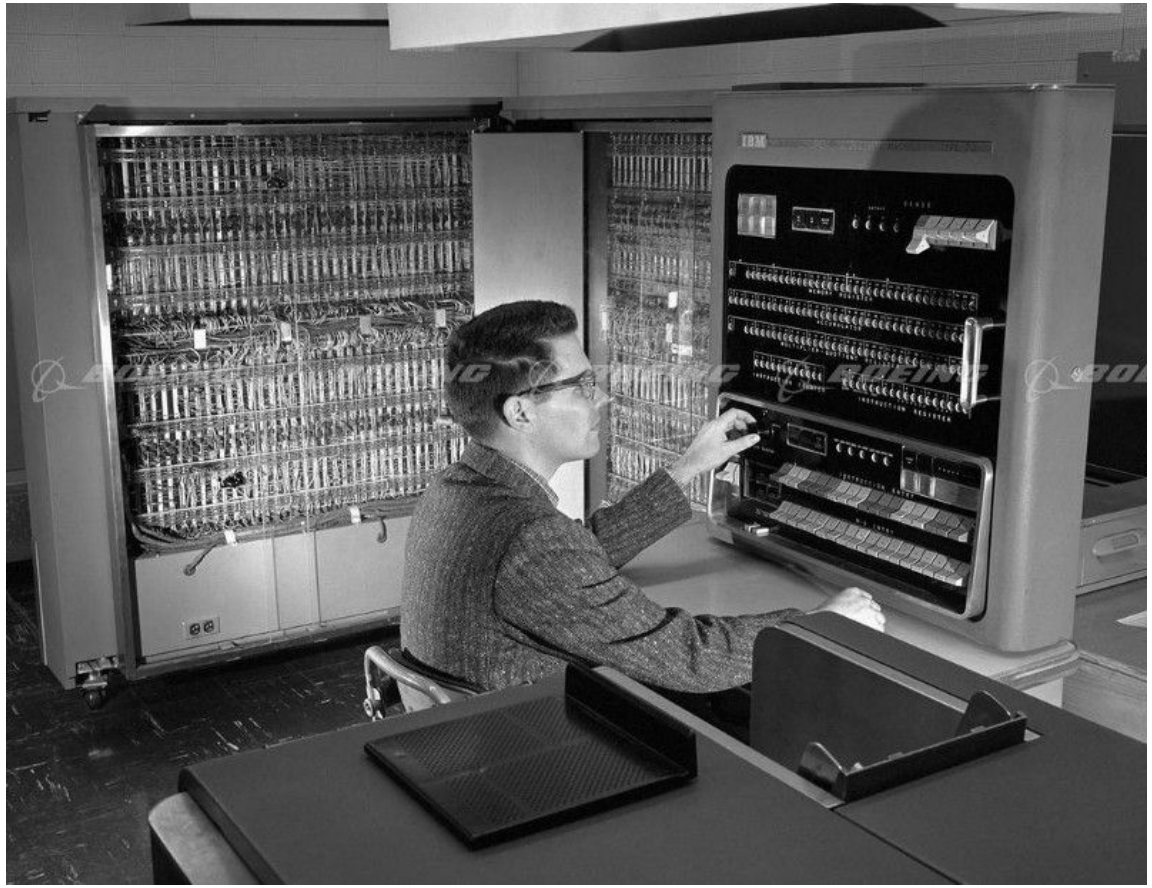



- In 1932, Georges Artsrouni created a “mechanical brain” to translate texts from one language to another by matching words with a paper tape, ignoring grammar rules
- In 1933, Peter Troyanskii proposed to use a dictionary along with a method of learning the grammar of a language based on Esperanto

# Machine Translation

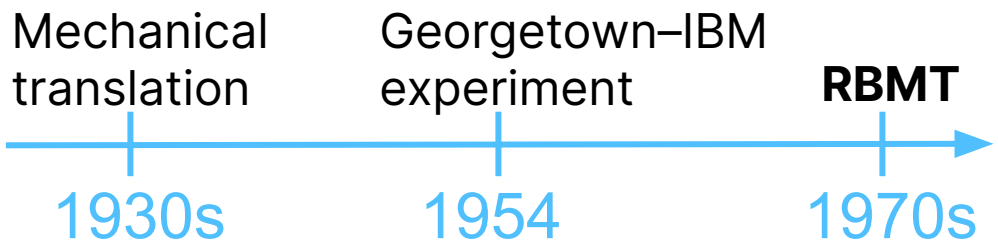



- Ru-En translation of 60 transliterated sentences
- The first public demonstration of machine translation capabilities
- The system consisted of six grammar rules and a private glossary of 250 lexical elements



IBM 701

# Machine Translation

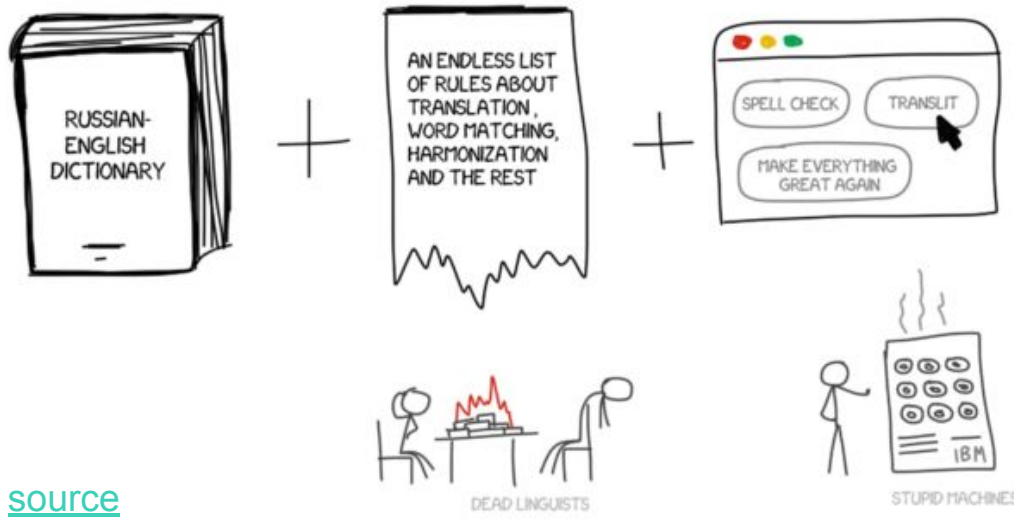



- Direct Machine Translation

  `give me a pen -> geben ich ein Kuli`
- Transfer-based Machine Translation

  `i bought an apple -> ich habe gekauft einen Apfel`
- Interlingual Machine Translation

  `i like cats ->` 😄 😍 🐈 `-> ich mag Katzen`




source

# Machine Translation

- morphological accuracy
- reproducibility
- subject area tuning

Mechanical translation — 1930s

Georgetown–IBM experiment — 1954

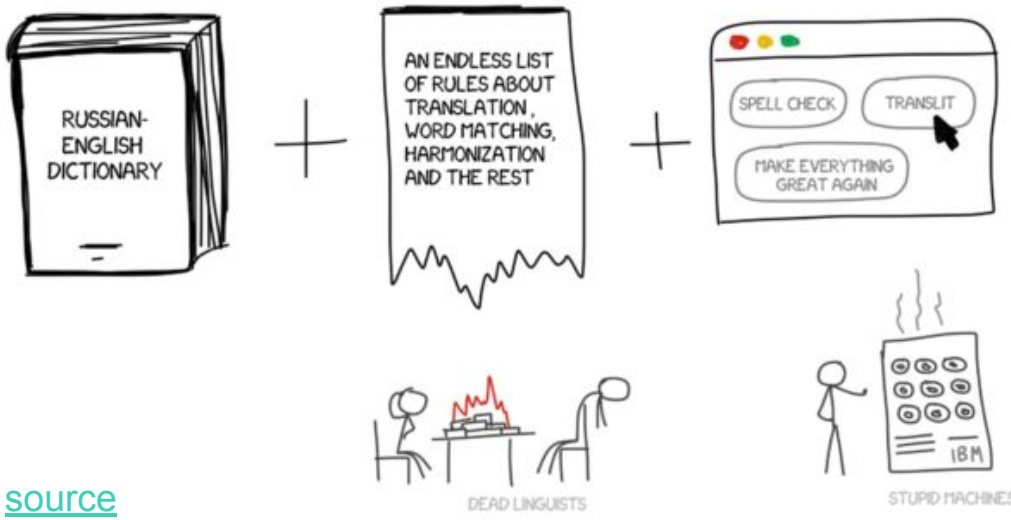
**RBMT** — 1970s

- Direct Machine Translation

  `give me a pen -> geben ich ein Kuli`
- Transfer-based Machine Translation

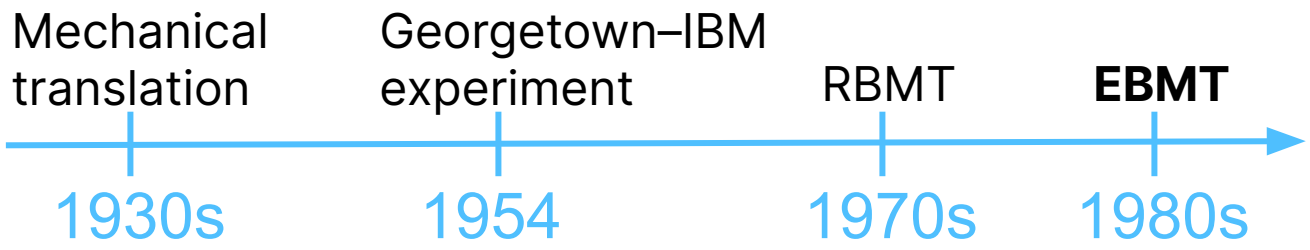
  `i bought an apple -> ich habe gekauft einen Apfel`
- Interlingual Machine Translation

  `i like cats ->` 😄 😍 🐈 `-> ich mag Katzen`




source

# Machine Translation



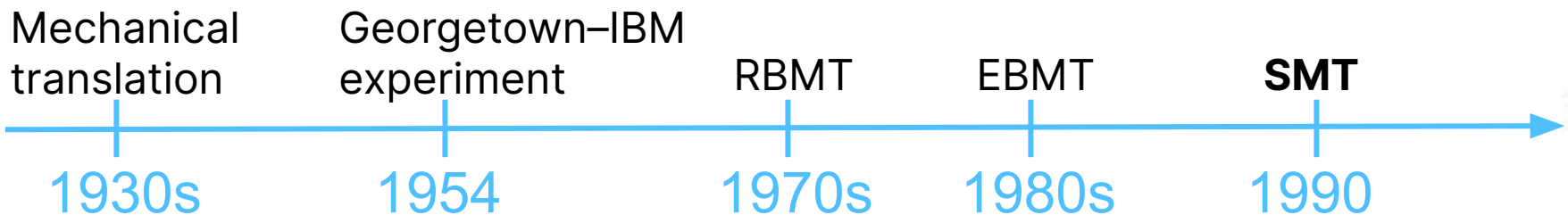

- Bilingual corpus with parallel texts
- Translation by analogy

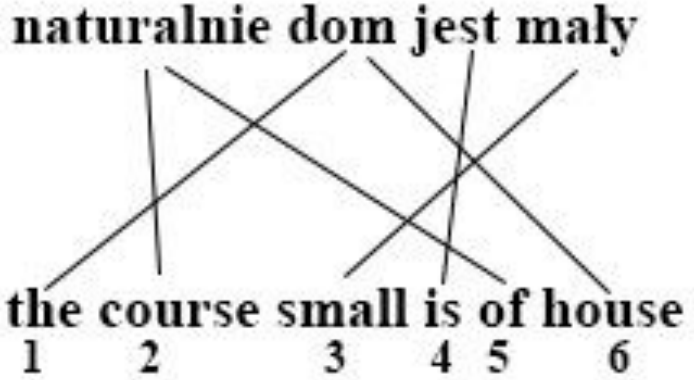
```
i like cats -> ich mag Katzen
i like dogs -> ich mag Hunde
```

- started from Japanese-English translation in 1984

Makoto Nagao

# Machine Translation



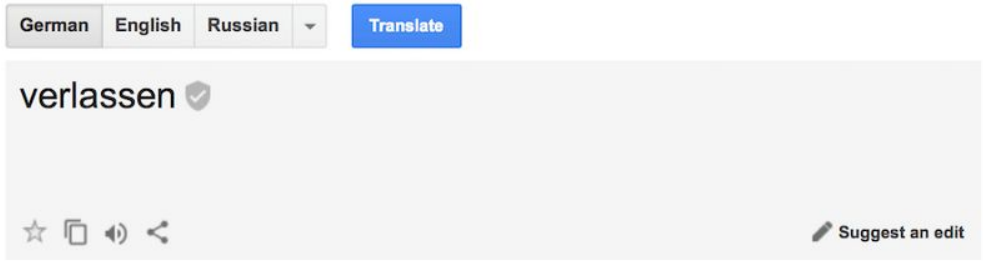



source

- The source text is partitioned into a set of fixed locations, then a sequence is created with the glossary, and words in target locations are rearranged to form a target sentence
- The statistical techniques for automatic glossary creation and arrangement of target word sequences
- Five statistical models (IBM alignment models) for the translation and algorithms for the parameters estimation the parameters
- No linguists needed, only larger corpora: the abstracts of the European Parliament and the United Nations Security Council meetings are available in the languages of all member countries

AIRI / Университет Сириус

# Machine Translation




Translations of **leave**

*verb*

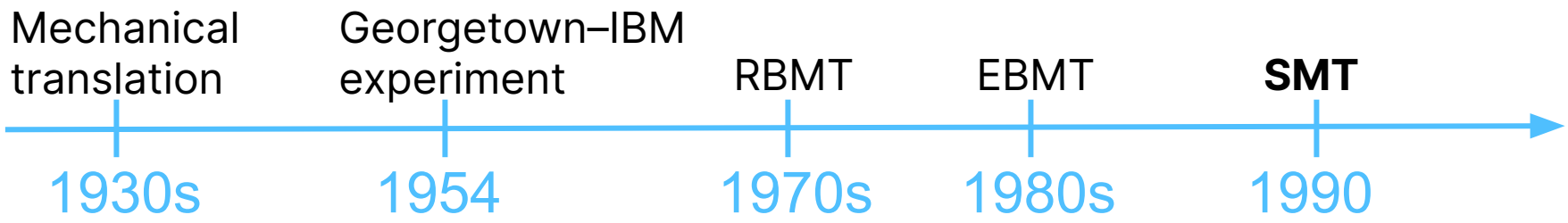
| | |
|---|---|
| verlassen | leave, abandon, exit, quit, forsake, desert |
| lassen | let, leave, allow, stop, let go, let be |
| hinterlassen | leave, leave behind, bequeath |
| abgeben | leave, dispense, deliver, submit, emit, give off |
| gehen | go, walk, leave, move, go down, quit |
| überlassen | leave, entrust, leave up to, leave with, intrust, let have |
| belassen | leave |
| abreisen | leave, depart, check out |
| stehen lassen | let stand, leave, leave behind, leave untouched, abandon |
| zurücklassen | leave, leave behind |
| austreten | escape, leave, withdraw, come out, resign, opt out |
| abfahren | leave, depart, go, ski down, wear, move off |
| liegen lassen | leave, leave out, leave lying around, leave behind |
| fortgehen | go away, leave, depart, quit |
| vergessen | forget, forget about |
| abgehen | leave, depart, come off, go off, exit, sail |
| abfliegen | depart, leave, take off, fly off, go |
| sein lassen | leave, stop, leave alone, drop |
| ausfahren | extend, come out, go for a ride, deliver |
| vermachen | bequeath, leave, will, make over |
| übrig bleiben | remain, leave, be left over |
| vererben | bequeath, pass on, leave, hand down |
| fortfahren | continue, proceed, go on, resume, leave, go away |
| sich entfernen | move away, leave, go away, recede, depart, move off |
| hinausfahren | drive out, go out, leave |

probabilities

reverse statistics

AIRI / Университет Сириус

# Machine Translation

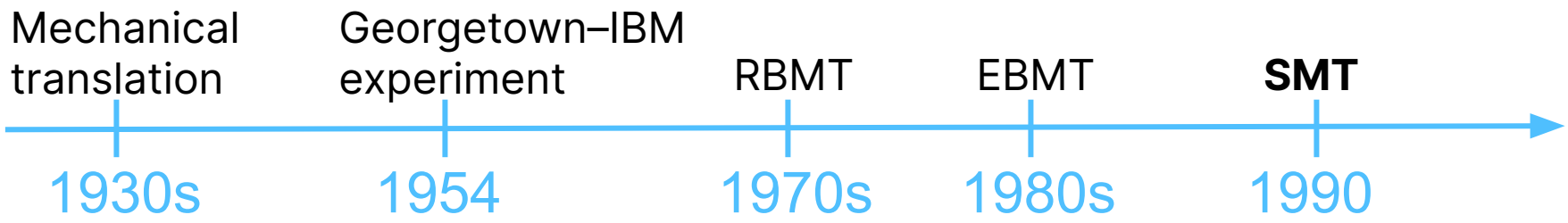



- Word-based
  - *the bag of words → the sentence word order → new words handling → word alignment with relative word order*
- Phrase-based
  - word-based techniques applied to n-grams, wider range of texts for learning

Frederick Jelinek

*“Every time I fire a linguist, the performance of the speech recognizer goes up.”*

# Machine Translation




- Word-based
  - *the bag of words → the sentence word order → new words handling → word alignment with relative word order*
- Phrase-based
  - word-based techniques applied to n-grams, wider range of texts for learning

2006

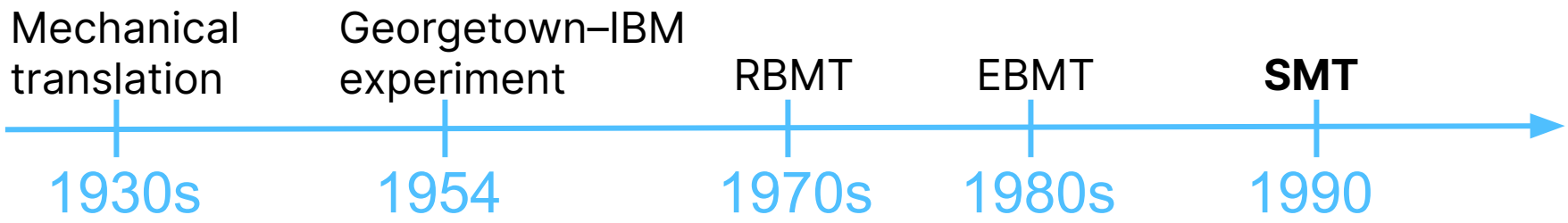
2016

Frederick Jelinek

*“Every time I fire a linguist, the performance of the speech recognizer goes up.”*

AIRI / Университет Сириус

# Machine Translation




- Word-based
  - *the bag of words → the sentence word order → new words handling → word alignment with relative word order*
- Phrase-based
  - word-based techniques applied to n-grams, wider range of texts for learning
- Syntax-based
  - careful syntax analysis combined with rules

2006

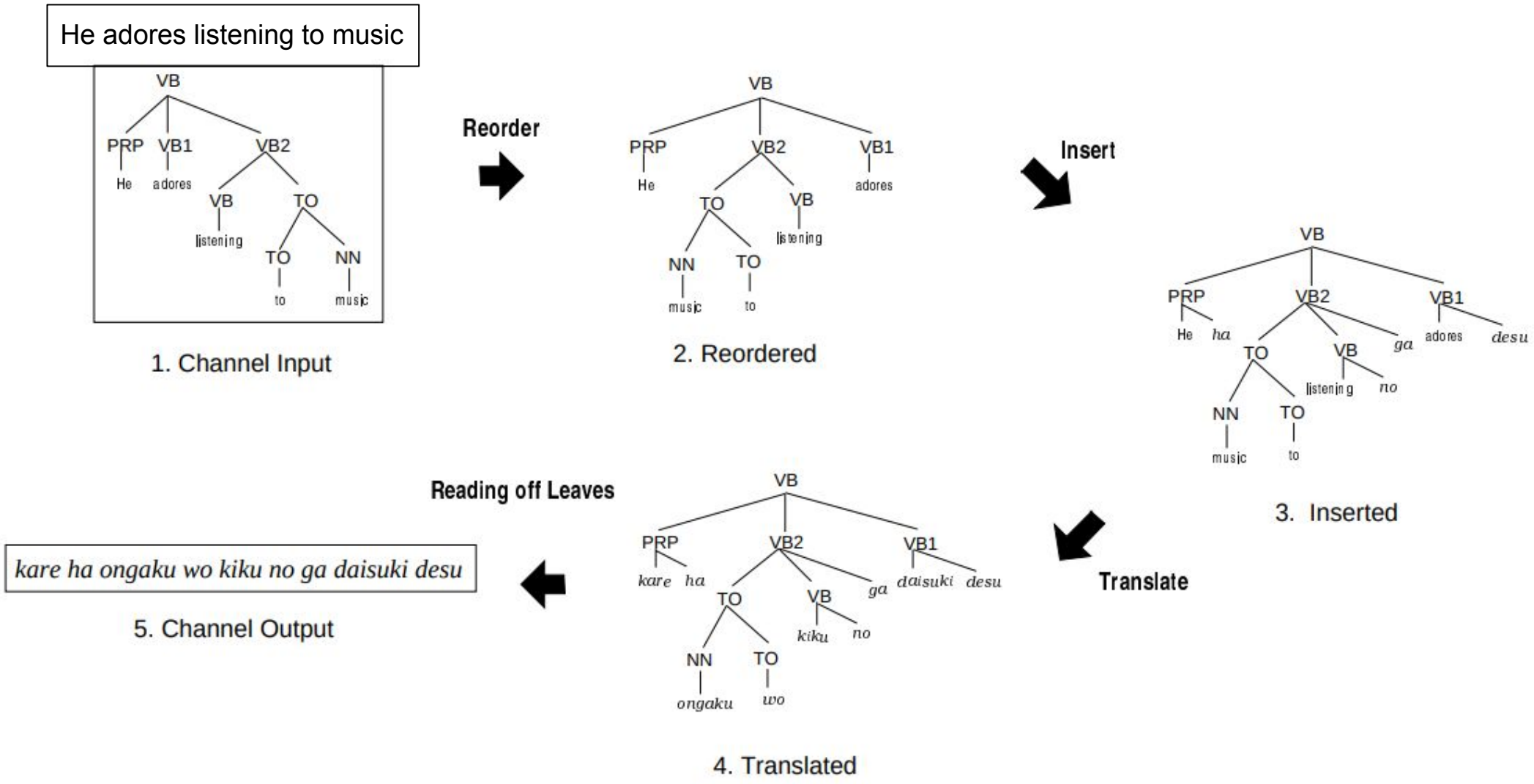

2016

Frederick Jelinek

*“Every time I fire a linguist, the performance of the speech recognizer goes up.”*

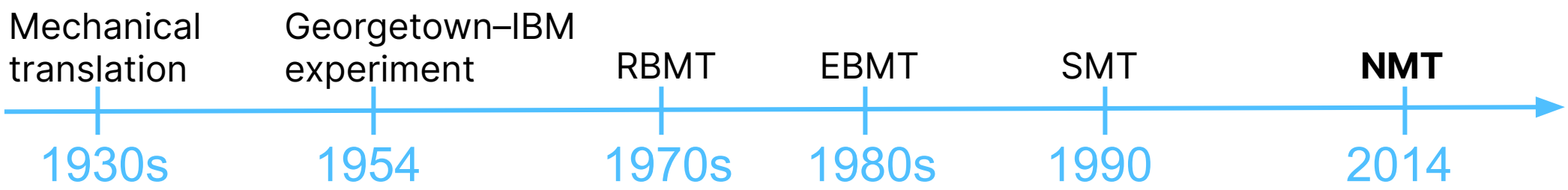
AIRI / Университет Сириус

# Machine Translation

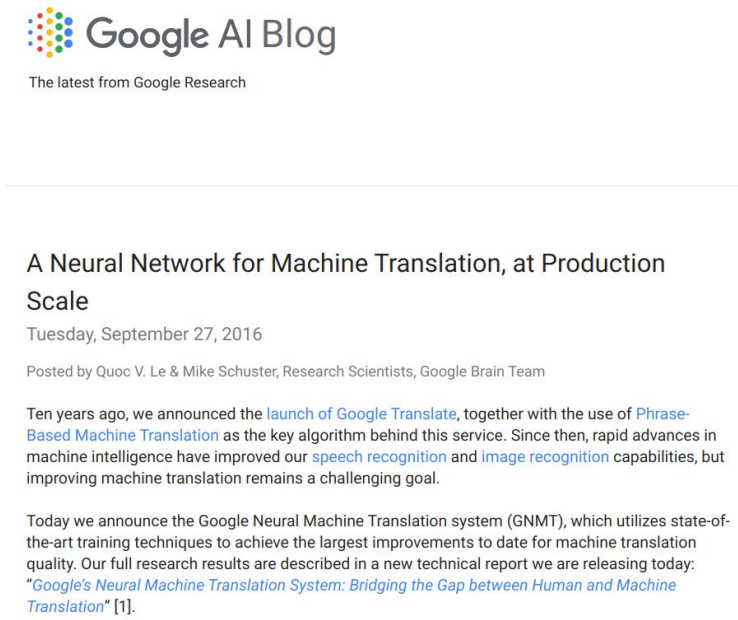



[Kenji Yamada, Kevin Knight, “A Syntax-based Statistical Translation Model”, 2001]

# Machine Translation




- RNNs: Continuous Translation Models
- Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation
- Google: RNN-based machine translation (GNMT)
- Yandex: combined neural + statistical MT + CatBoost to choose the best variant


Google AI Blog

The latest from Google Research

A Neural Network for Machine Translation, at Production Scale

Tuesday, September 27, 2016

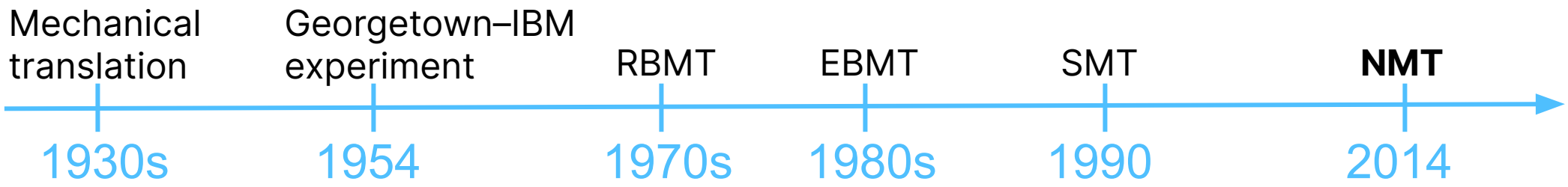
Posted by Quoc V. Le & Mike Schuster, Research Scientists, Google Brain Team

Ten years ago, we announced the launch of Google Translate, together with the use of Phrase-Based Machine Translation as the key algorithm behind this service. Since then, rapid advances in machine intelligence have improved our speech recognition and image recognition capabilities, but improving machine translation remains a challenging goal.

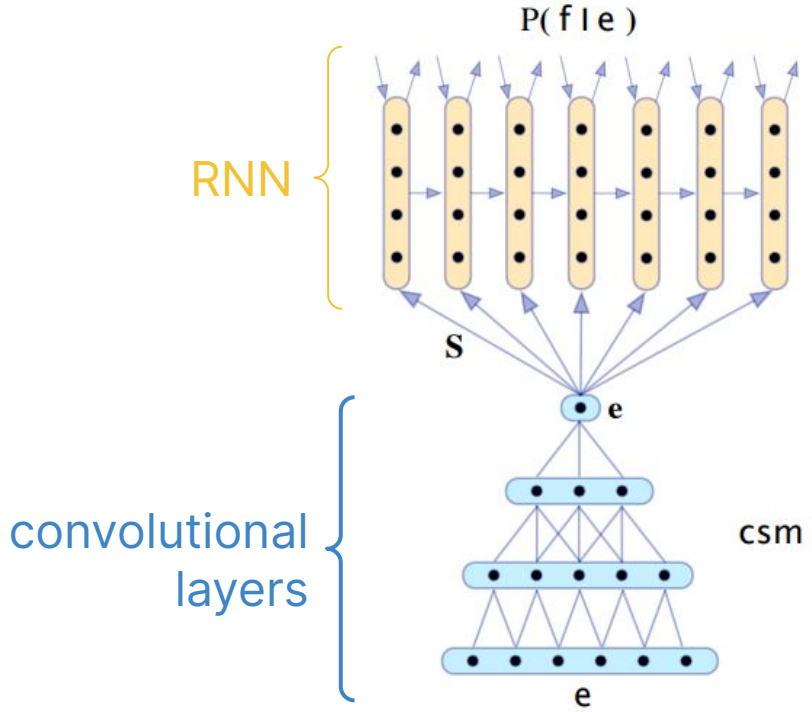
Today we announce the Google Neural Machine Translation system (GNMT), which utilizes state-of-the-art training techniques to achieve the largest improvements to date for machine translation quality. Our full research results are described in a new technical report we are releasing today: "*Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*" [1].

AIRI / Университет Сириус

# Machine Translation

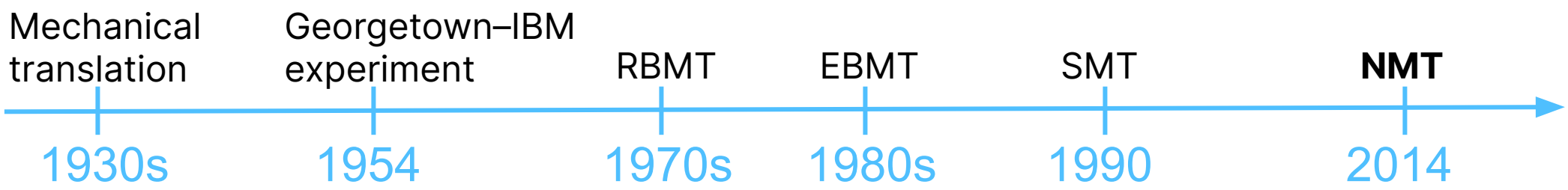



- RNNs
  - Continuous Translation Models

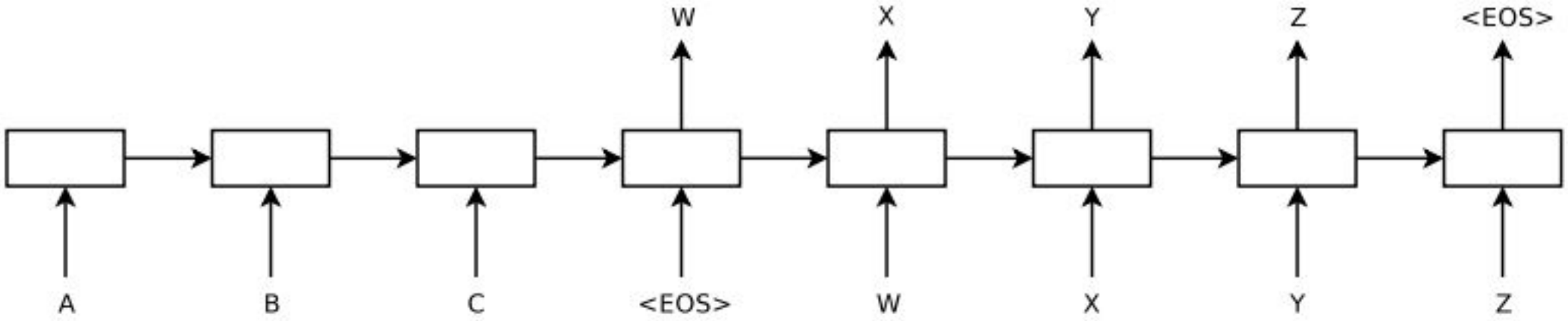



AIRI / Университет Сириус

# Machine Translation

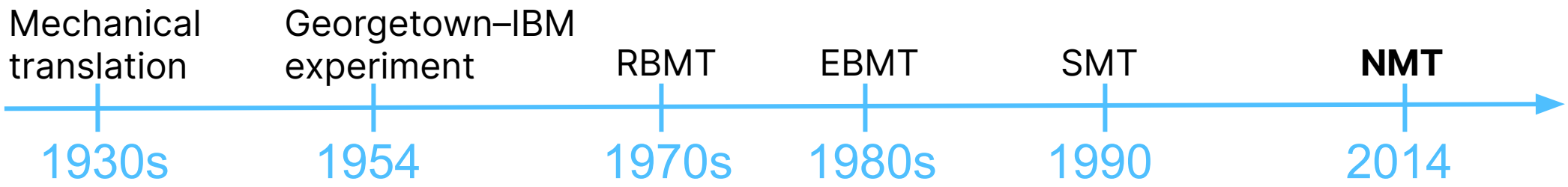



- RNNs
  - Continuous Translation Models
  - LSTM Seq2seq

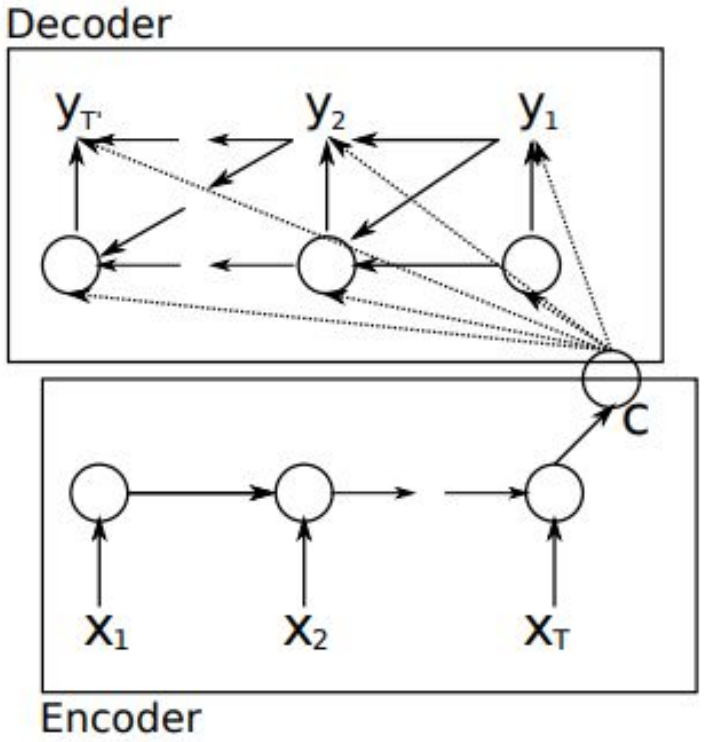



AIRI / Университет Сириус

# Machine Translation

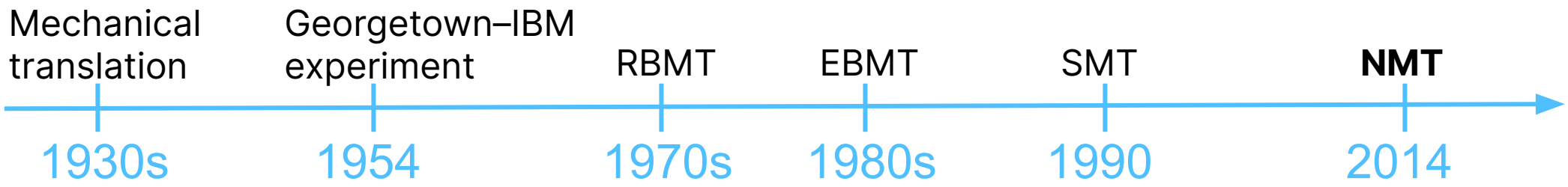



- RNNs
  - Continuous Translation Models
  - LSTM Seq2seq
  - Encoder-Decoder RNNs

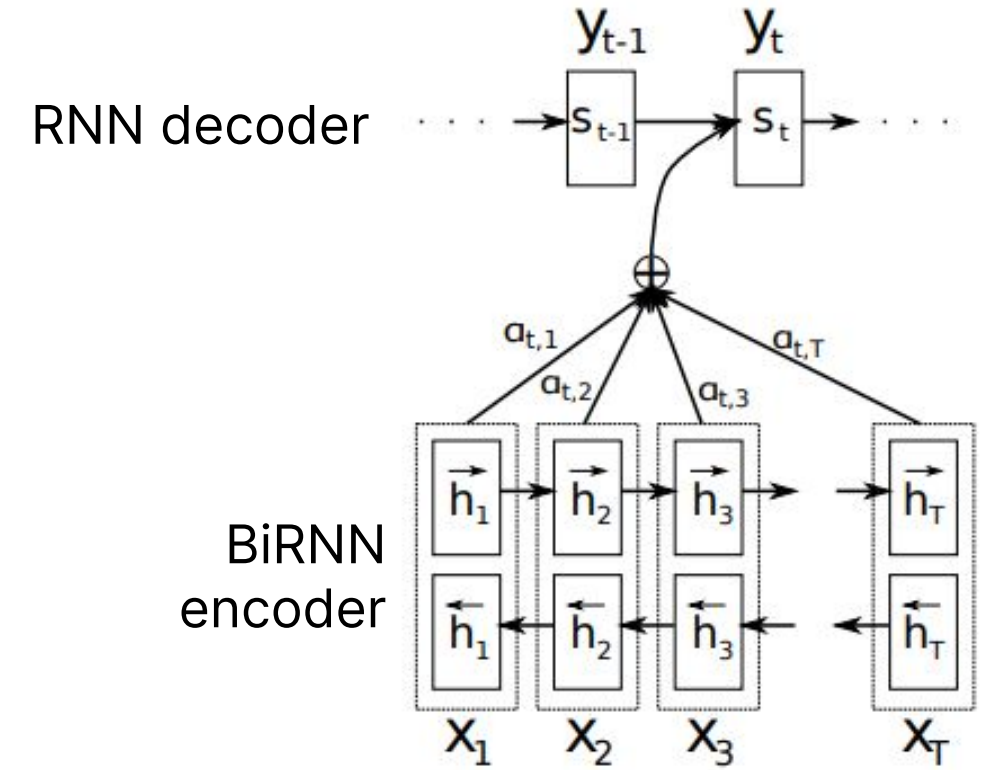



AIRI / Университет Сириус

# Machine Translation

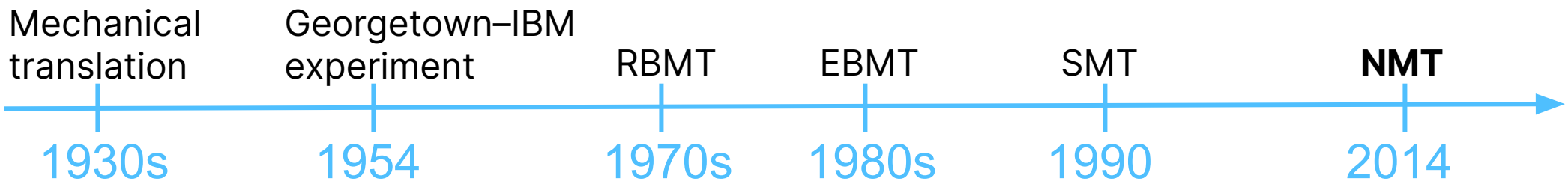



- RNNs
  - Continuous Translation Models
  - LSTM Seq2seq
  - Encoder-Decoder RNNs
  - Jointly Learning to Align and Translate

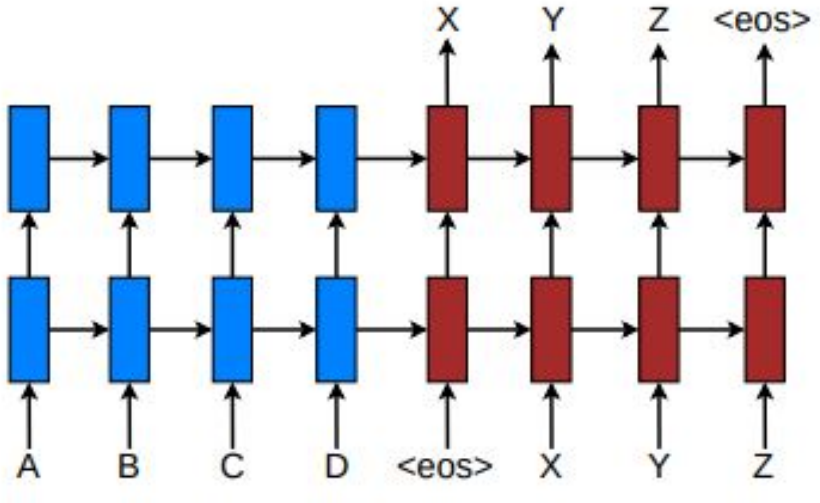



AIRI / Университет Сириус

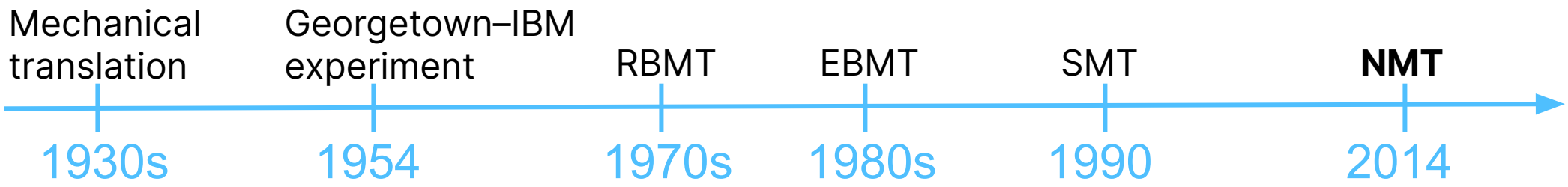
# Machine Translation



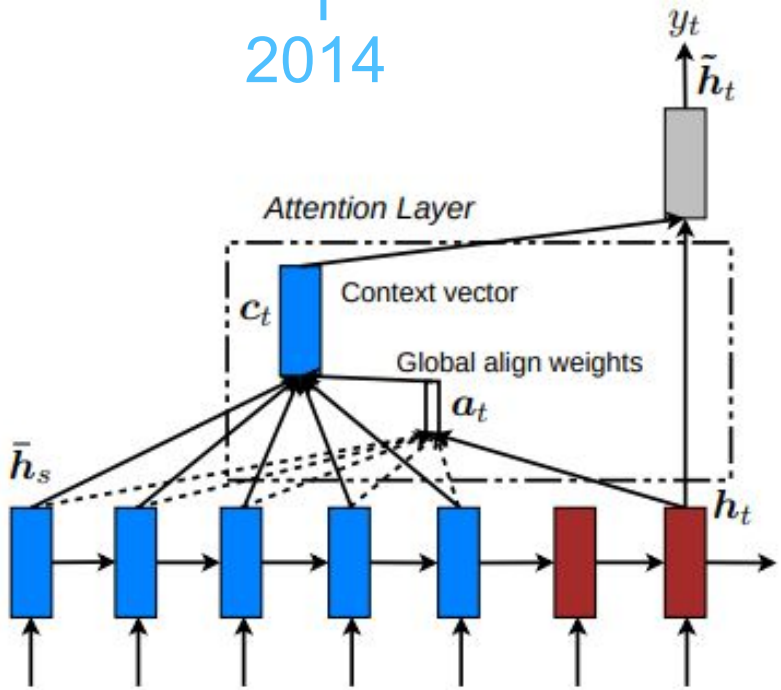

- RNNs
  - Continuous Translation Models
  - LSTM Seq2seq
  - Encoder-Decoder RNNs
  - Jointly Learning to Align and Translate
  - Attention-based NMT



# Machine Translation

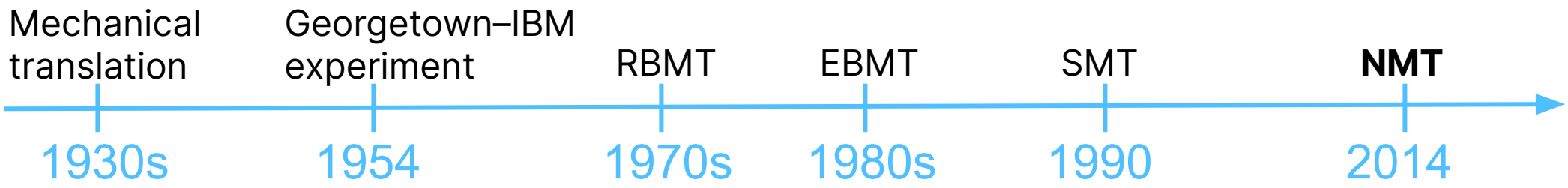



- RNNs
  - Continuous Translation Models
  - LSTM Seq2seq
  - Encoder-Decoder RNNs
  - Jointly Learning to Align and Translate
  - Attention-based NMT

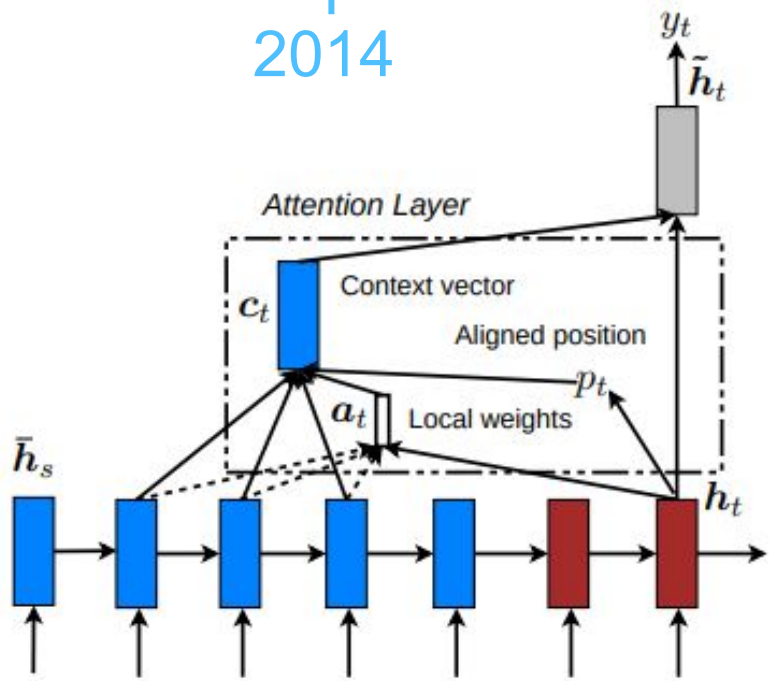



Global attention

# Machine Translation

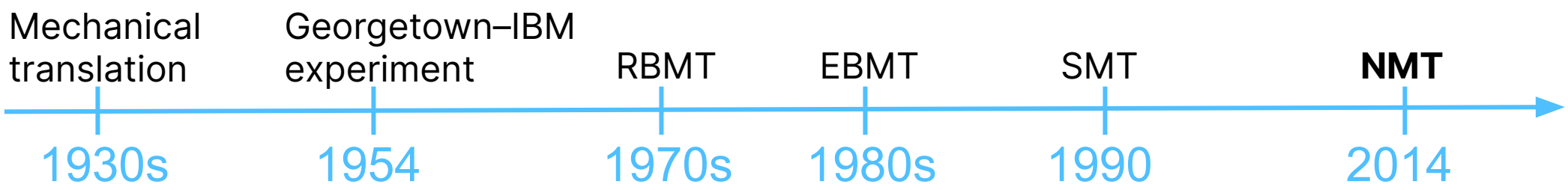



- RNNs
  - Continuous Translation Models
  - LSTM Seq2seq
  - Encoder-Decoder RNNs
  - Jointly Learning to Align and Translate
  - Attention-based NMT

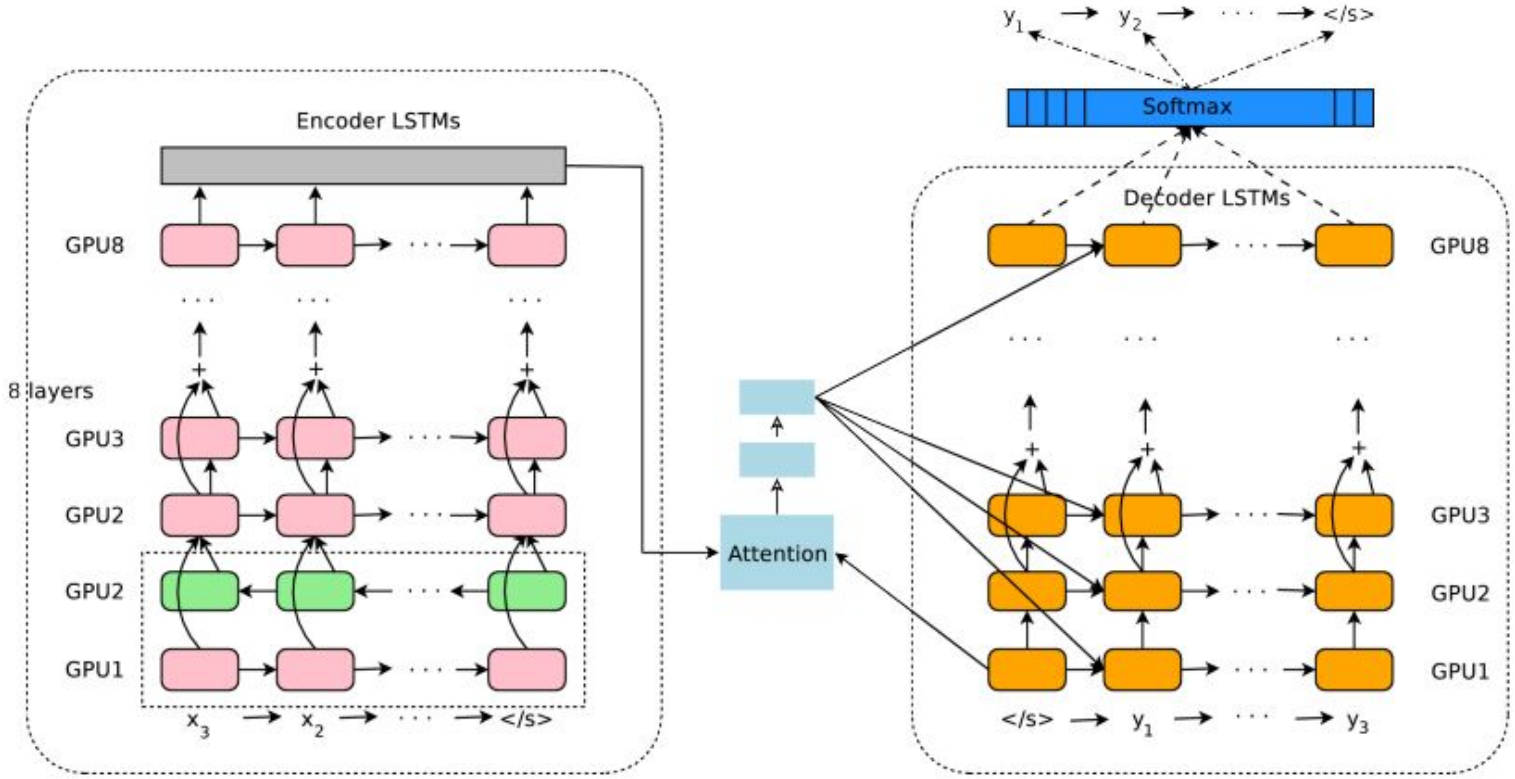



Local attention

# Machine Translation




- RNNs
  - Continuous Translation Models
  - LSTM Seq2seq
  - Encoder-Decoder RNNs
  - Jointly Learning to Align and Translate
  - Attention-based NMT
  - GNMT



# 02

# Generative Models

# Generative Models: recap

**Context (human-written):** In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

---

**GPT-2:** The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

# Generative Models

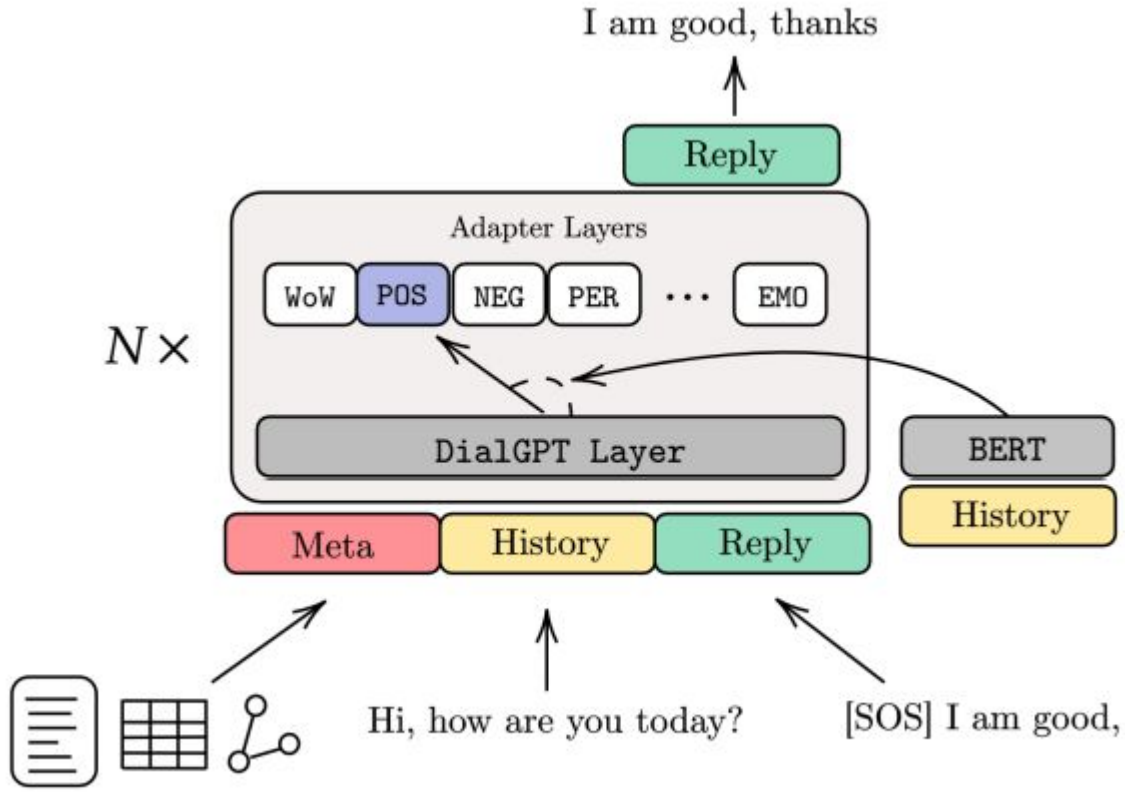
- 2018-2020: GPT (117M) → GPT-2 (1.5B) → GPT-3 (**175B**)
- 2019: Multi-Modality to Multi-Modality Multitask Mega-transformer (M6) (10B)
- 2022: OPT-175B - trained on publicly available data

**Applications:**

- creation of articles, poetry, stories, news reports
- (controllable) dialogue generation, machine translation, Q&A, summarization
- product description generation, visual question answering, community question answering,  text-guided image generation
- natural language generation, the ability to detect biases and hate speech

# The Adapter-Bot: All-In-One Controllable Conversational Model

- DialoGPT (117M-762M)
- dialogue skills like emphatic response, weather information, movie recommendation implemented via adapters
- adapters are trained independently
- knowledge sources: text, tables, graphs
- skills triggered with dialogue manager/manually

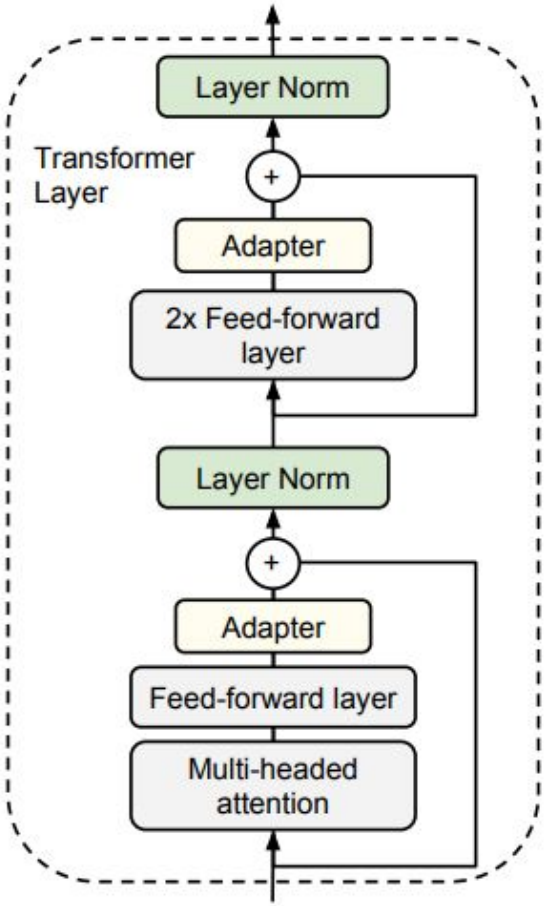



source

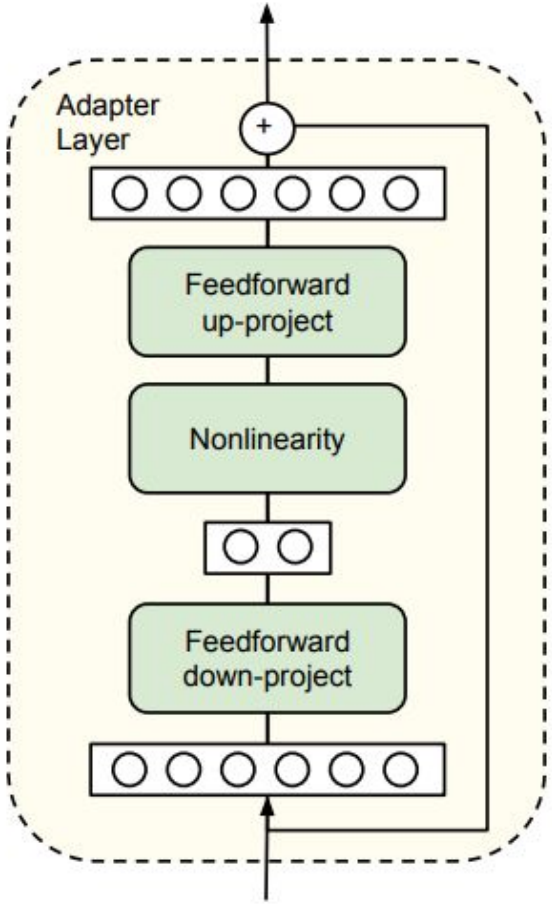
# Adapter module

Steers the output distribution of a pre-trained model without modifying the original weights.

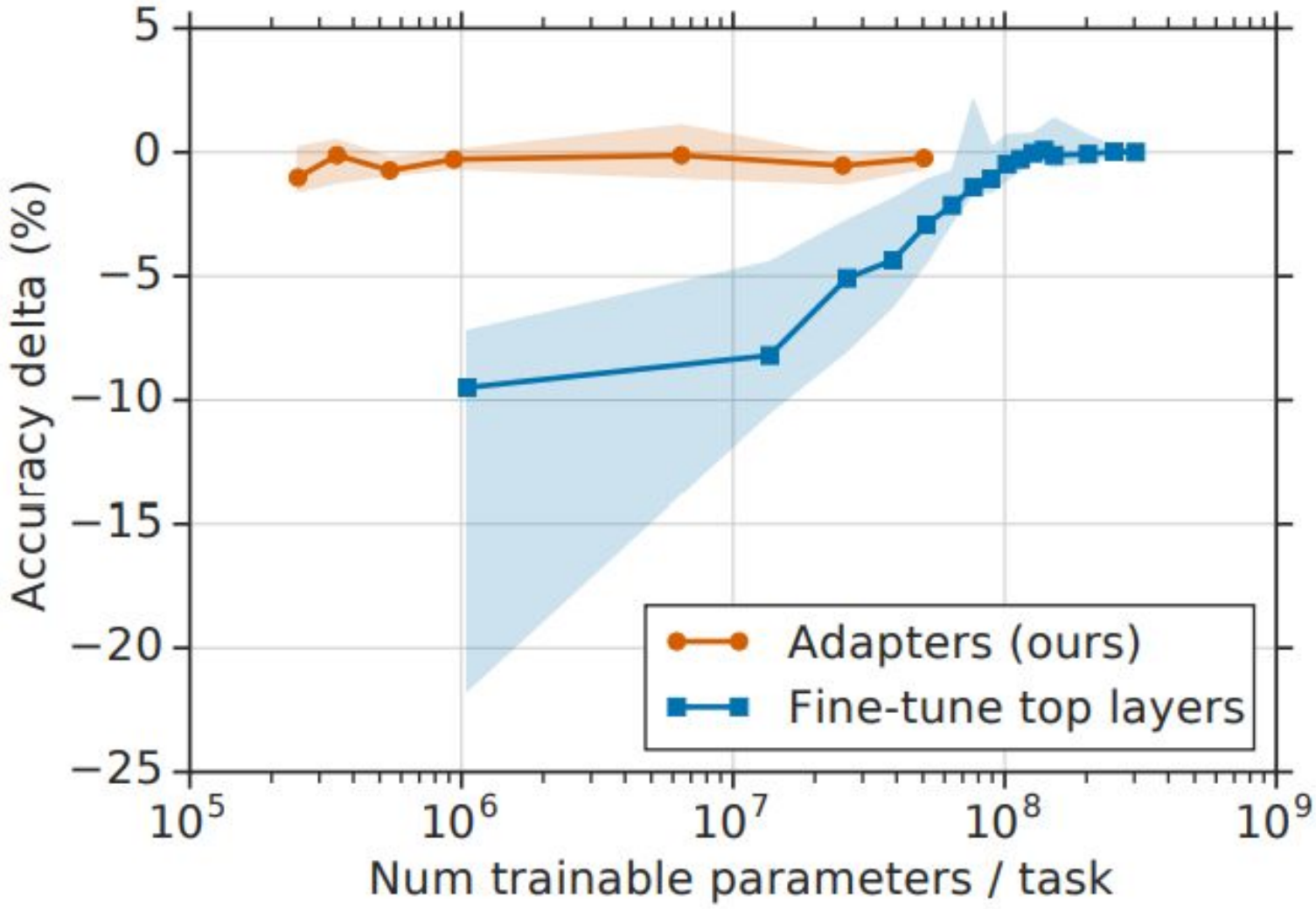
- a bottleneck with few parameters relative to the attention and feedforward layers in the original model and a skip-connection
- green layers are trained on the downstream data, this includes the adapter, the layer normalization parameters, and the final classification layer




source



# Adapter module




source

# Retrieval-Enhanced Transformer (Retro)

- Retrieving from a database with trillions of tokens
- Key-value database
- Retrieval as a way to augment input examples at the granularity of small chunks of tokens

# Retrieval-Enhanced Transformer (Retro)

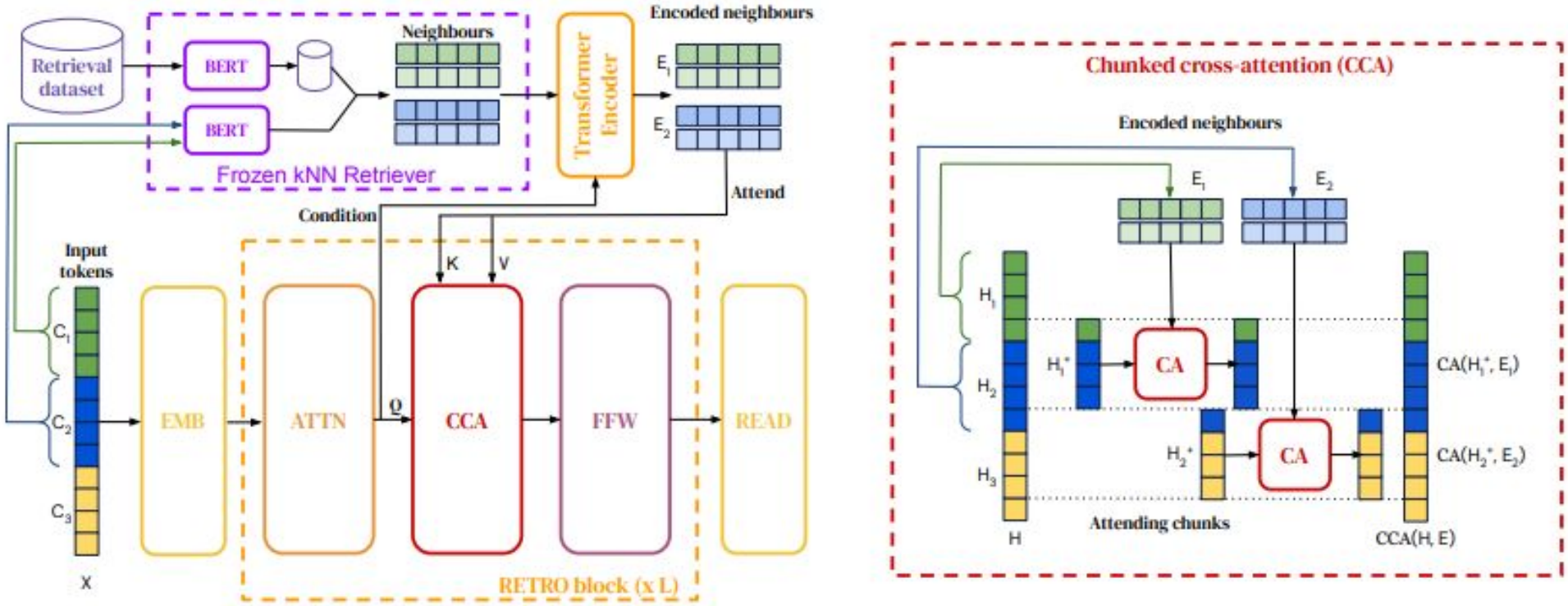
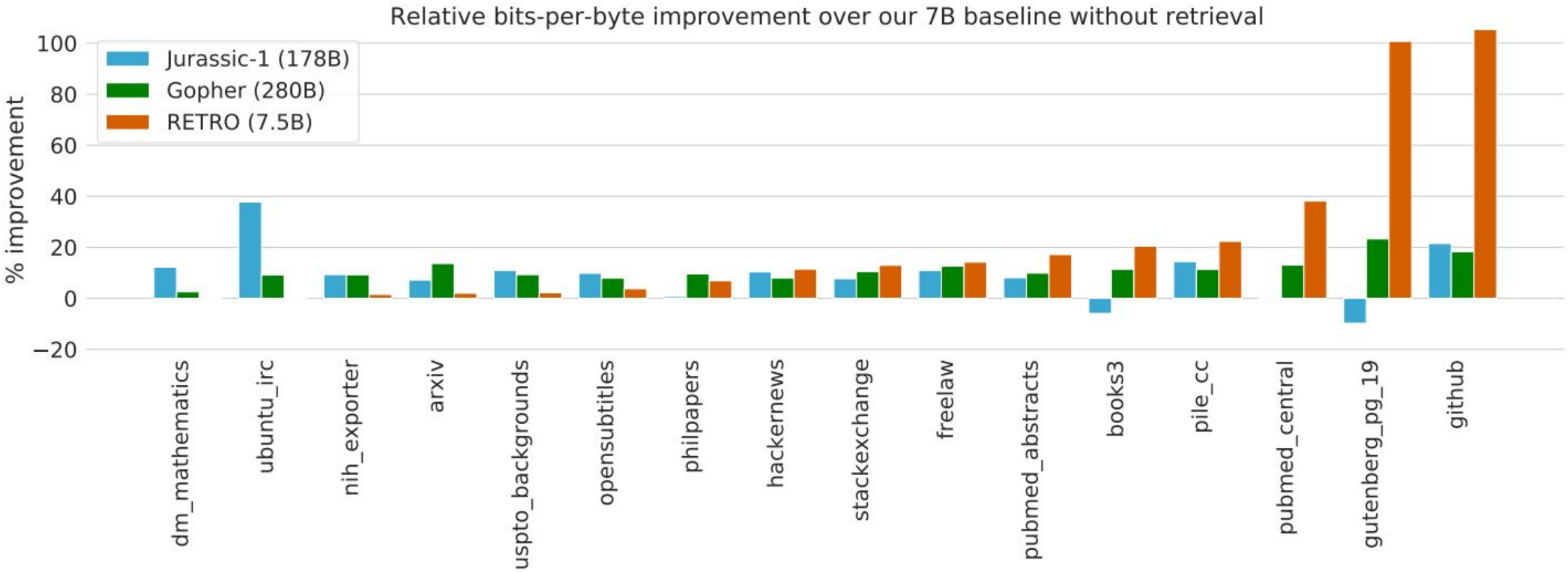



Figure 2 | **RETRO architecture.** *Left:* simplified version where a sequence of length $n = 12$ is split into $l = 3$ chunks of size $m = 4$. For each chunk, we retrieve $k = 2$ neighbours of $r = 5$ tokens each. The retrieval pathway is shown on top. *Right:* Details of the interactions in the CCA operator. Causality is maintained as neighbours of the first chunk only affect the last token of the first chunk and tokens from the second chunk.

source

# Retrieval-Enhanced Transformer (Retro)




source

# PaLM: Scaling Language Modeling with Pathways

- exploring impact of scale on few-shot in-context learning


**6 Evaluation** 11

6.1 English NLP tasks . . . 11
6.2 BIG-bench . . . 14
6.3 Reasoning . . . 19
6.4 Code Tasks . . . 21
6.5 Translation . . . 27
6.6 Multilingual Natural Language Generation . . . 29
6.7 Multilingual Question Answering . . . 32
6.8 Analysis . . . 32


- Pathways training system to train a single model across thousands or tens of thousands of accelerator chips (TPUs) in a highly efficient manner

# PaLM: prompting

## Explaining a Joke

I will explain these jokes:
(1) The problem with kleptomaniacs is that they always take things literally.
Explanation: This joke is wordplay. Someone who "takes things literally" is someone who doesn't fully understand social cues and context, which is a negative trait. But the definition of kleptomania is someone who literally takes things.
(2) Always borrow money from a pessimist. They'll never expect it back.
Explanation: Most people expect you to pay them back when you borrow money, however a pessimist is someone who always assumes the worst, so if you borrow money from them, they will expect that you won't pay them back anyways.

**Input:** I was supposed to start writing the paper at 5:00 PM. But then I started playing with this cool new language model for 10 minutes. 10 minutes later, it's suddenly 9:30 PM!
**Model Output:** This joke is about how time flies when you're having fun. The person was supposed to start writing a paper at 5:00 PM, but they started playing with a language model instead. They thought they were only playing with it for 10 minutes, but it was actually 4.5 hours.

**Input:** I think I understand why me and my partner are such a good match. I get excited when I smell food and become super alert when the doorbell rings. My girlfriend Shelley becomes ecstatic when she sees a person who she just saw 5 hours ago.
**Model Output:** This joke is about dogs. Dogs get excited when they smell food and when the doorbell rings. The girlfriend is excited to see someone she just saw 5 hours ago, which is a trait of dogs.

**Input:** I was going to fly to visit my family on April 6th. My mom said "Oh great, your step dad's poetry reading is that night!" So now I'm flying in on April 7th.
**Model Output:** The joke is that the speaker's mother is trying to get them to go to their step dad's poetry reading, but the speaker doesn't want to go, so they are changing their flight to the day after the poetry reading.

**Input:** I tried 10,000 random restarts of my neural network, but I was accused of overfitting. I guess no good seed goes unpunished.
**Model Output:** This joke is a pun. A neural network is a computer program that can learn from data. A "seed" is a number that is used to initialize a random number generator. A "good seed" is a number that produces a good random number generator. The phrase "no good deed goes unpunished" means that if you do something good, you will be punished for it.

**Input:** Reading "War and Peace" doesn't scare me. I once had to review the CL for my intern's whole project on his final day!
**Model Output:** 'War and Peace' is a very long book, so reading it is a daunting task. However, the speaker is saying they had to review the code for their intern's project on their final day, which is even more daunting.

**Input:** What's the difference between a zebra and an umbrella? One is a striped animal related to horses, another is a device you use to stop the rain from falling on you.
**Model Output:** This joke is an anti-joke. The joke is that the answer is obvious, and the joke is that you were expecting a funny answer.

# PaLM: prompting

## Logical Inference

I will explain how to make logical inferences:
(1) John wakes to the sound of his children giggling outside the door. They bust in and say "Happy Father's Day!" with a plate full of food. 2022 has been a great year. Can we infer that John was born before the inauguration of Barack Obama?
Answer: We can infer that John is a father because of "his children". We can infer that the current year is 2022. Barack Obama was inaugurated in 2009, so it has been about 13 years since Obama was inaugurated. John's children are old enough to speak, so if John were born after Obama's inauguration, he would have been younger than 10 when he had children. So, we can infer that John was born before Obama's inauguration. The answer is "yes".
(2) Sam sees a piano in the subway stations. Laughing, his friends prod him to play, saying "Remember how amazing your rendition of Chopsticks was at the Christmas party?" Can we infer that John majored in piano at Juilliard?
Answer: Juilliard is a top school for music, so someone who majored in piano at Juilliard would be an expert pianist. "Chopsticks" is a very simple song to play on the piano. His friends were therefore likely being sarcastic when they told him to play, which would imply that Sam was not an expert piano player. So it is unlikely that Sam majored in piano at Juilliard. The answer is "no".

**Input:** When I found out my grandma was in the hospital I felt a particular color. When someone cut me off in traffic I felt a different color. What is the most likely color I would see if I combined these two colors?
**Model Output:** The color that I felt when I found out my grandma was in the hospital was likely a sad color, like blue. The color that I felt when someone cut me off in traffic was likely an angry color, like red. If I combined these two colors, I would likely see purple. The answer is "purple".

**Input:** Michael is at that really famous museum in France looking at its most famous painting. However, the artist who made this painting just makes Michael think of his favorite cartoon character from his childhood. What was the country of origin of the thing that the cartoon character usually holds in his hand?
**Model Output:** The most famous painting in the Louvre is the Mona Lisa. The artist who made the Mona Lisa is Leonardo da Vinci. Leonardo da Vinci is also the name of the main character in the cartoon Teenage Mutant Ninja Turtles. Leonardo da Vinci is from Italy. The thing that Leonardo da Vinci usually holds in his hand is a katana. The country of origin of the katana is Japan. The answer is "Japan".

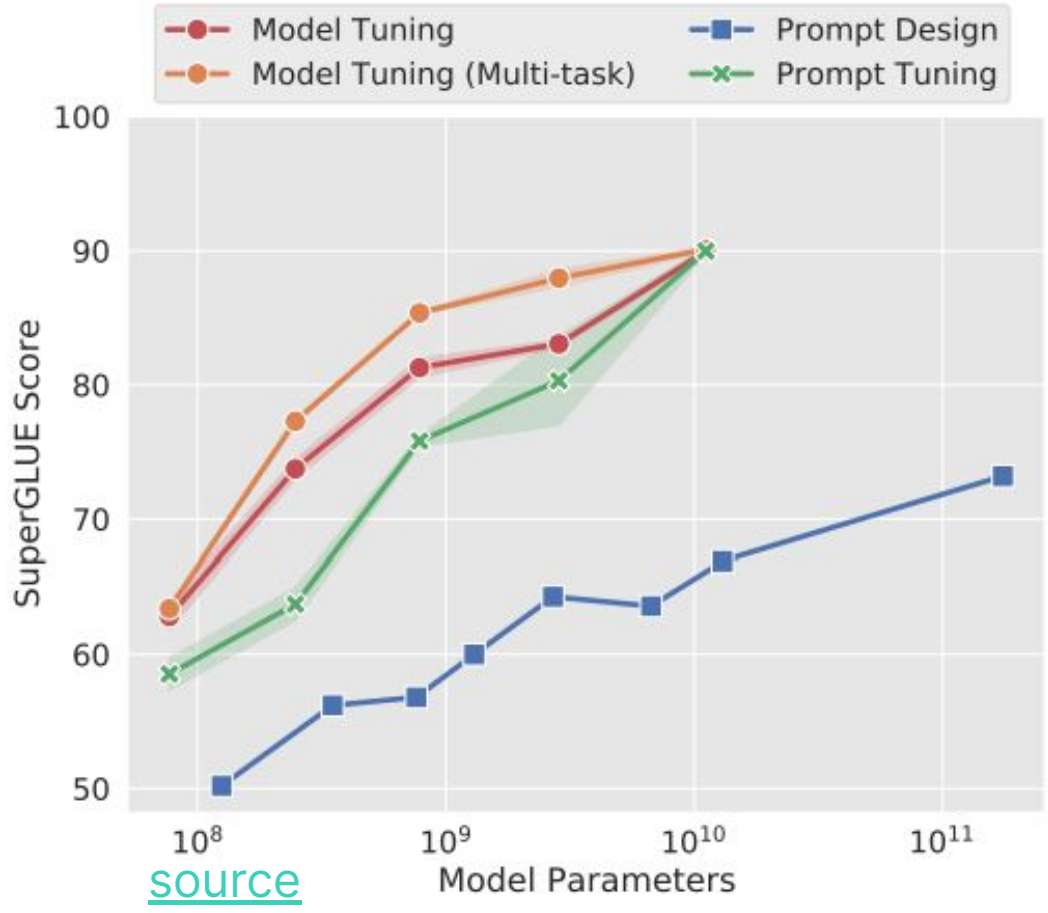
**Input:** Guido was writing a Python program. He keeps getting confused by variable scope and looking it up on StackOverflow. Is it possible that the Guido in the prompt refers to Guido van Rossum?
**Model Output:** Guido van Rossum is the creator of Python. He would not need to look up variable scope on StackOverflow. The answer is "no".

**Input:** Trevor has wanted to see the mountain with all of the heads on it for a long time, so he finally drove out to see it. What is the capital of the state that is directly east of the state that Trevor is currently in?
**Model Output:** The mountain with all of the heads on it is Mount Rushmore. Mount Rushmore is in South Dakota. The state directly east of South Dakota is Minnesota. The capital of Minnesota is St. Paul. The answer is "St. Paul".

**Input:** Jennifer looked out her window and sees a really cool cloud below her. She unbuckles her seatbelt and heads to the bathroom. Is Jennifer probably traveling more than 300 miles per hour relative to the earth?
**Model Output:** 300 miles per hour is about 480 km/h. This is about the speed of a commercial airplane. Clouds are usually below airplanes, so Jennifer is probably on an airplane. The answer is "yes".

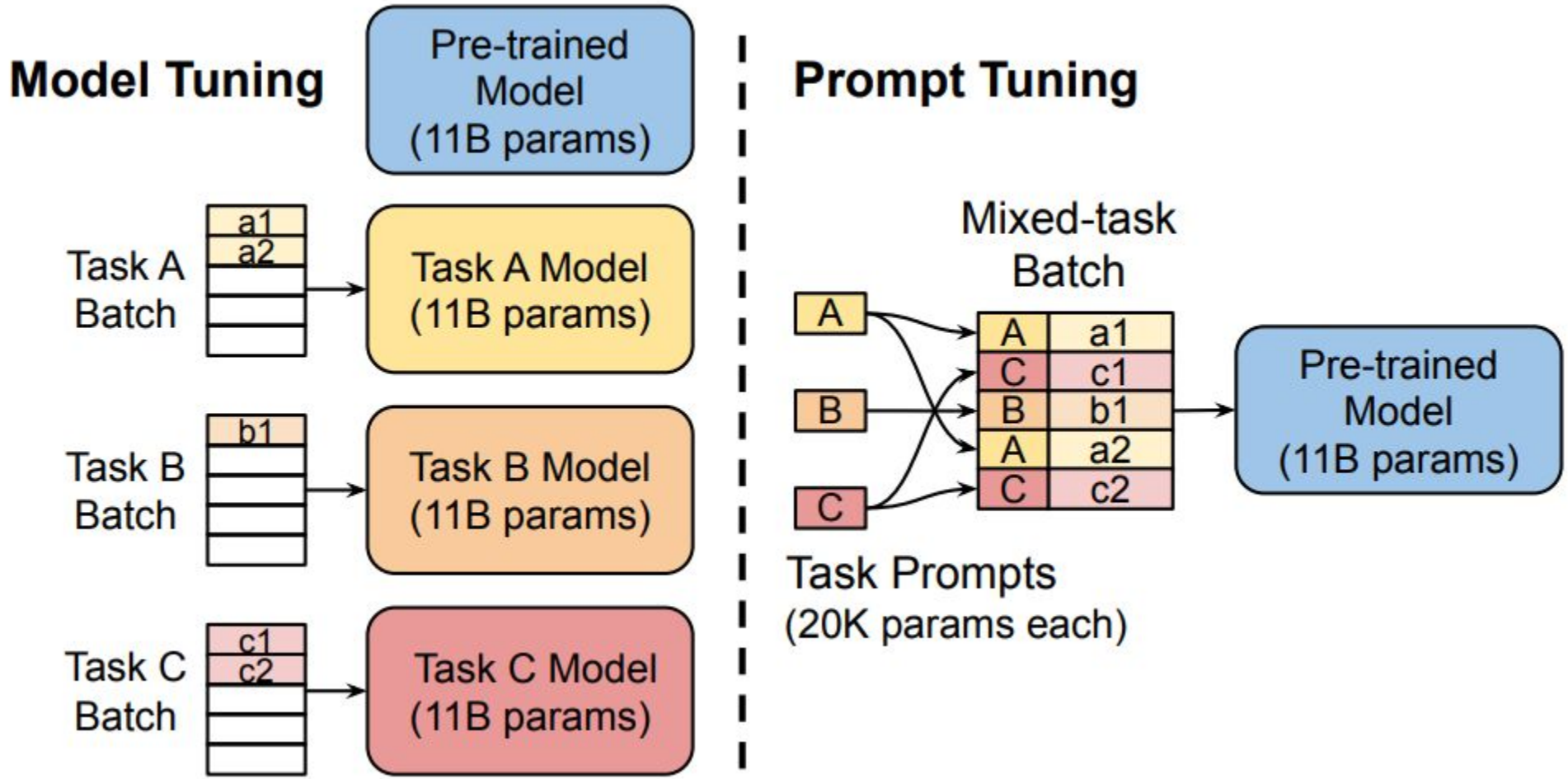
# The Power of Scale for Parameter-Efficient **Prompt Tuning**

- “soft prompts” help to condition frozen language models to perform downstream tasks
- prompt tuning can match the fine-tuned model performance with scale
- conditioning a frozen model with soft prompts enables efficient prompt ensembling




source

# Prompt Tuning



# Prompt Tuning

| Dataset | Domain | Model | Prompt | Δ |
|---|---|---|---|---|
| SQuAD | Wiki | 94.9 ±0.2 | 94.8 ±0.1 | −0.1 |
| TextbookQA | Book | 54.3 ±3.7 | **66.8** ±2.9 | +12.5 |
| BioASQ | Bio | 77.9 ±0.4 | **79.1** ±0.3 | +1.2 |
| RACE | Exam | 59.8 ±0.6 | **60.7** ±0.5 | +0.9 |
| RE | Wiki | 88.4 ±0.1 | **88.8** ±0.2 | +0.4 |
| DuoRC | Movie | **68.9** ±0.7 | 67.7 ±1.1 | −1.2 |
| DROP | Wiki | **68.9** ±1.7 | 67.1 ±1.9 | −1.8 |

Table 1: F1 mean and stddev for models trained on SQuAD and evaluated on out-of-domain datasets from the MRQA 2019 shared task. Prompt tuning tends to give stronger zero-shot performance than model tuning, especially on datasets with large domain shifts like TextbookQA.

| Train | Eval | Tuning | Accuracy | F1 |
|---|---|---|---|---|
| QQP | MRPC | Model | 73.1 ±0.9 | 81.2 ±2.1 |
| | | Prompt | **76.3** ±0.1 | **84.3** ±0.3 |
| MRPC | QQP | Model | 74.9 ±1.3 | **70.9** ±1.2 |
| | | Prompt | **75.4** ±0.8 | 69.7 ±0.3 |

Table 2: Mean and stddev of zero-shot domain transfer between two paraphrase detection tasks.

| Dataset | Metric | Average | Best | Ensemble |
|---|---|---|---|---|
| BoolQ | acc. | 91.1 | 91.3 | **91.7** |
| CB | acc./F1 | 99.3 / 99.0 | 100.00 / 100.00 | **100.0 / 100.0** |
| COPA | acc. | 98.8 | 100.0 | **100.0** |
| MultiRC | EM/F1$_a$ | 65.7 / 88.7 | 66.3 / 89.0 | **67.1 / 89.4** |
| ReCoRD | EM/F1 | 92.7 / 93.4 | 92.9 / 93.5 | **93.2 / 93.9** |
| RTE | acc. | 92.6 | **93.5** | **93.5** |
| WiC | acc. | 76.2 | 76.6 | **77.4** |
| WSC | acc. | 95.8 | **96.2** | **96.2** |
| SuperGLUE (dev) | | 90.5 | 91.0 | **91.3** |

Table 3: Performance of a five-prompt ensemble built from a single frozen T5-XXL model exceeds both the average and the best among the five prompts.

# Artificial Intelligence Research Institute

airi.net

- airi_research_institute
- AIRI Institute
- AIRI Institute
- AIRI_inst
- artificial-intelligence-research-institute