20 Years of WMT

Philipp Koehn

8 November 2025

Origin Story

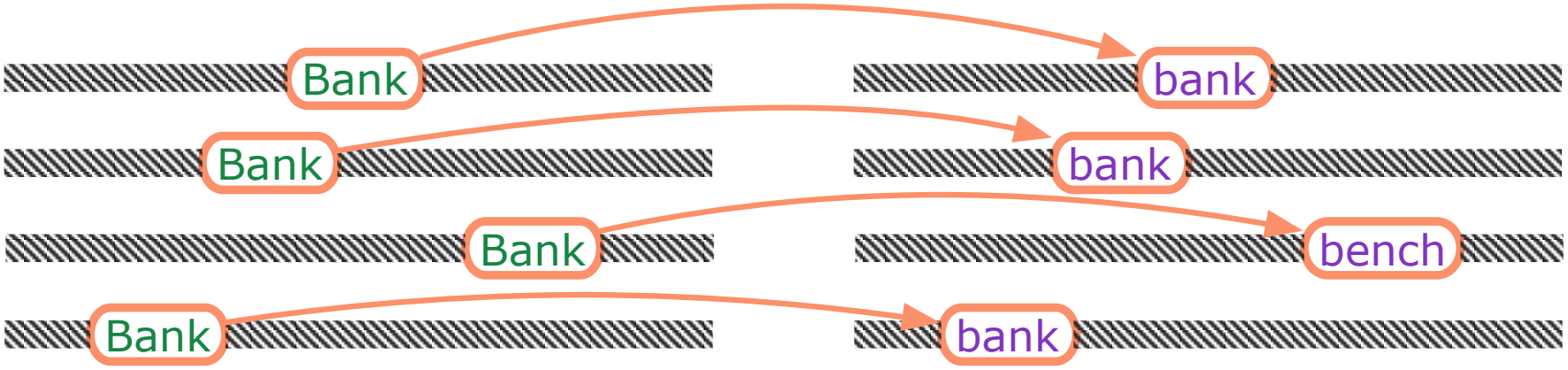
Vision 2004: an open evaluation campaign

First Event: 2005

- “Building and using parallel texts: data-driven MT and beyond”
- Merged with another workshop proposal
- Automatic evaluation on Europarl with BLEU scores

- Invited talk by Franz Och (Google): “Statistical Machine Translation: The Fabulous Present and Future”

Data-Driven Machine Translation



- Early battle of methods at first WMT (2006)
  - rule-based (Systran)
  - syntactically informed phrasal statistical (Microsoft)
  - inversion transduction grammars (Valencia)
  - n-gram translation model (UPC Barcelona)
  - hierarchical phrase models (NTT)

- Spanish, German, French $\leftrightarrow$ English

Progress in the First Decade

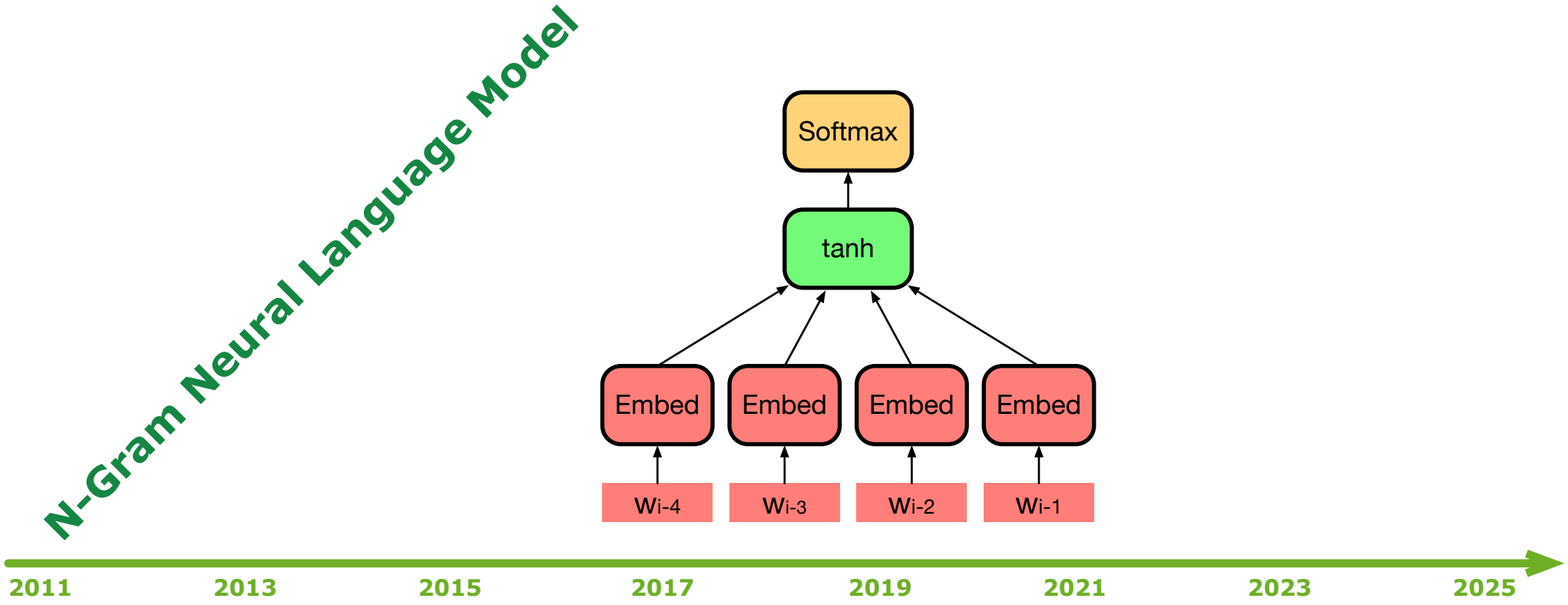
Early Large Language Models

Lang	Lines (B)	Tokens (B)	Bytes
en	59.13	975.63	5.14 TiB
de	3.87	51.93	317.46 GiB
fr	3.04	49.31	273.96 GiB
ru	1.79	21.41	220.62 GiB
cs	0.47	5.79	34.67 GiB
hi	0.01	0.28	3.39 GiB

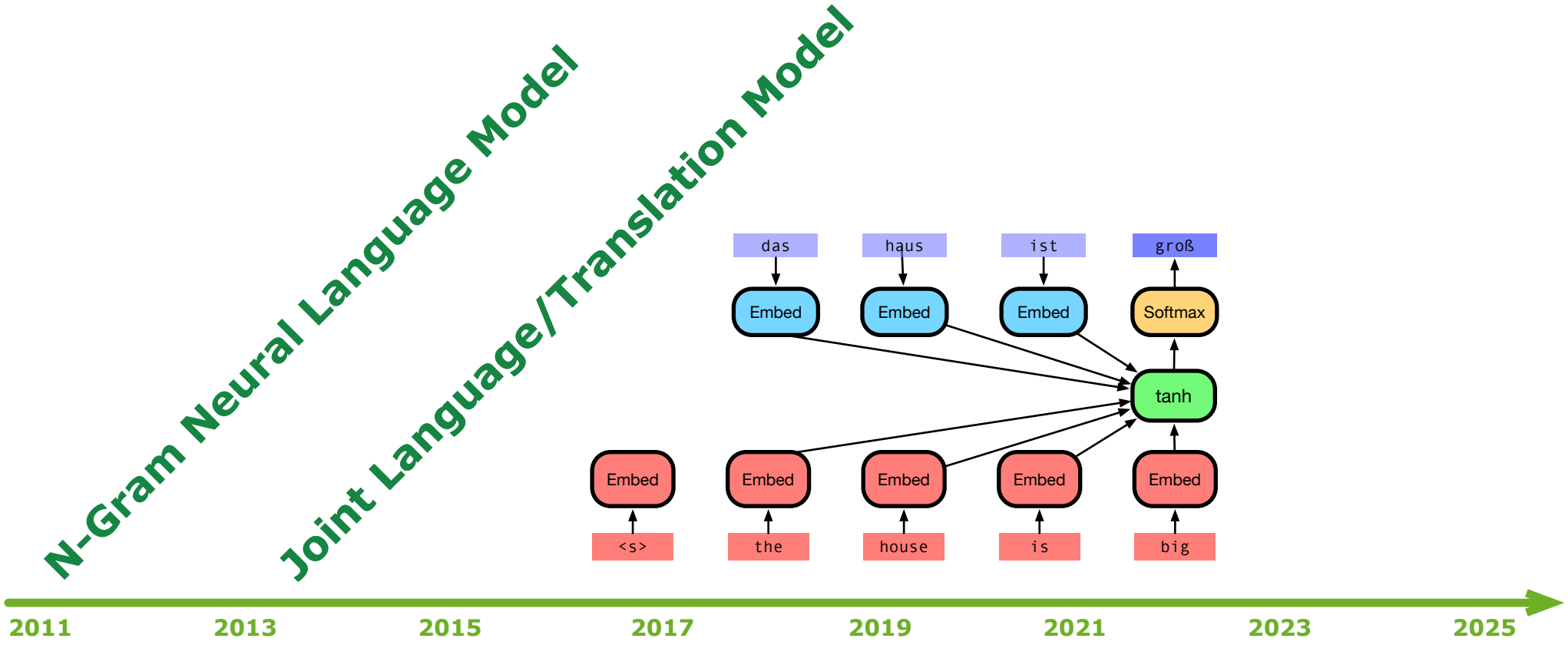
Table 5: Size of huge language model training data

- WMT14: LM trained on 975.63 billion tokens (English), from CommonCrawl
- Compare GPT2: 8 billion tokens

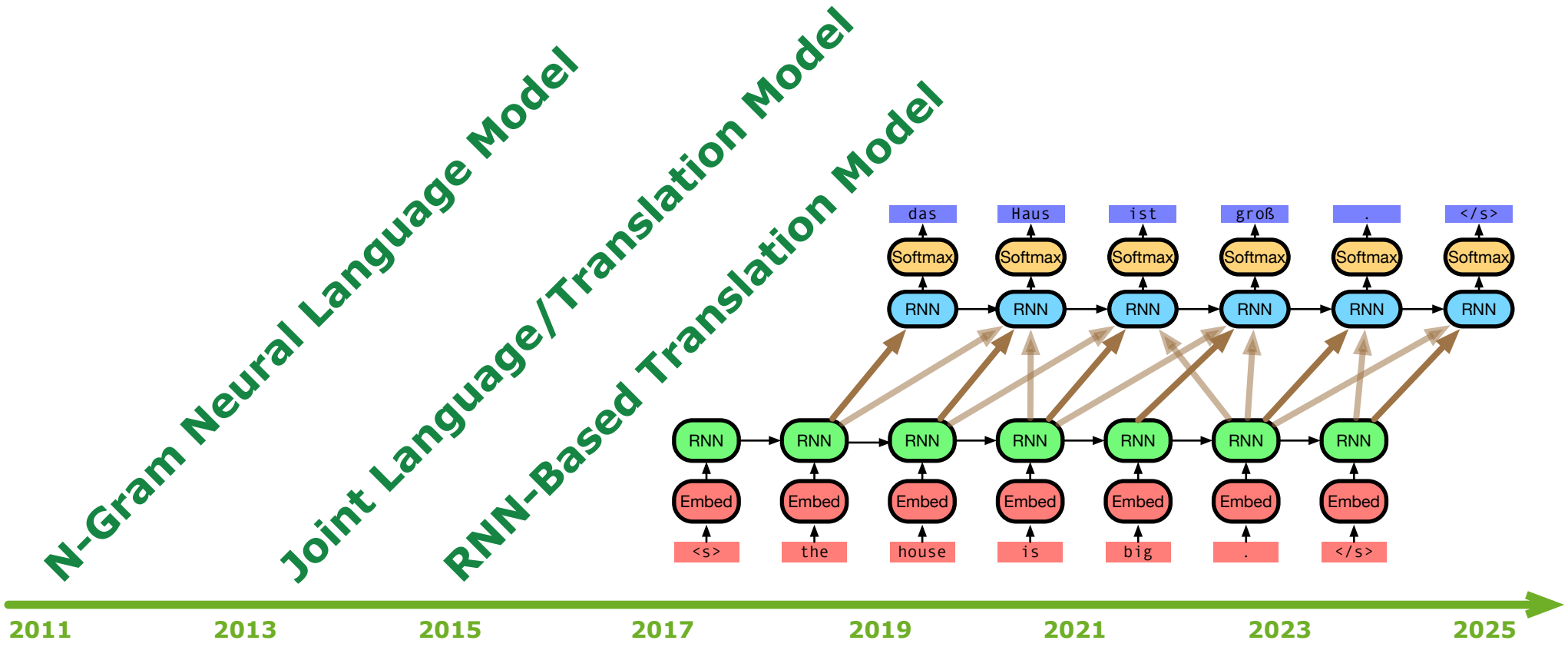
Neural Advances



Neural Advances



Neural Advances



Transformer

- Transformer model is a MT model, it uses WMT 2014 test sets

- A gradual evolution
  - n-gram neural language models (used in WMT since 2011)
  - recurrent neural network language models
  - RNN-based sequence-to-sequence model with attention
  - dot-product attention
  - attention replaces recurrent neural network

- Jargon term “decoder-only” LLM betrays this history

LLM

- Oh, yeah, monolingual data matters

- LLMs emerged as alternative to NMT (WMT 2023)
  — essentially all submissions today

- These were never “unsupervised” with MT as “emergent feature”
  — definitely not now

- Key innovations for LLMs come out of MT research
  - transformer model
  - byte pair encoding
  - data synthesis (back translation)

lessons learned

Obsession with Evaluation

- Human evaluation
  - struggling with inter-annotator agreement
  - mix of metrics
    - adequacy/fluency
    - ranking
    - absolute scoring on 100-point scale that has an odd name
    - MQM / ESL
  - consolidate scores, compute pairwise comparison and statistical significance?

- Automatic evaluation: are metrics breaking with increased quality?
  - BLEU
  - TER, METEOR, chrF
  - COMET and other trained metrics

- New test sets every year, manually created, no overlap with training

Open Science

- Open source software
  (Moses, PyTorch, etc.)

- Openly available data resources
  (WMT, OPUS, CommonCrawl, etc.)

OPUS

COMMON CRAWL

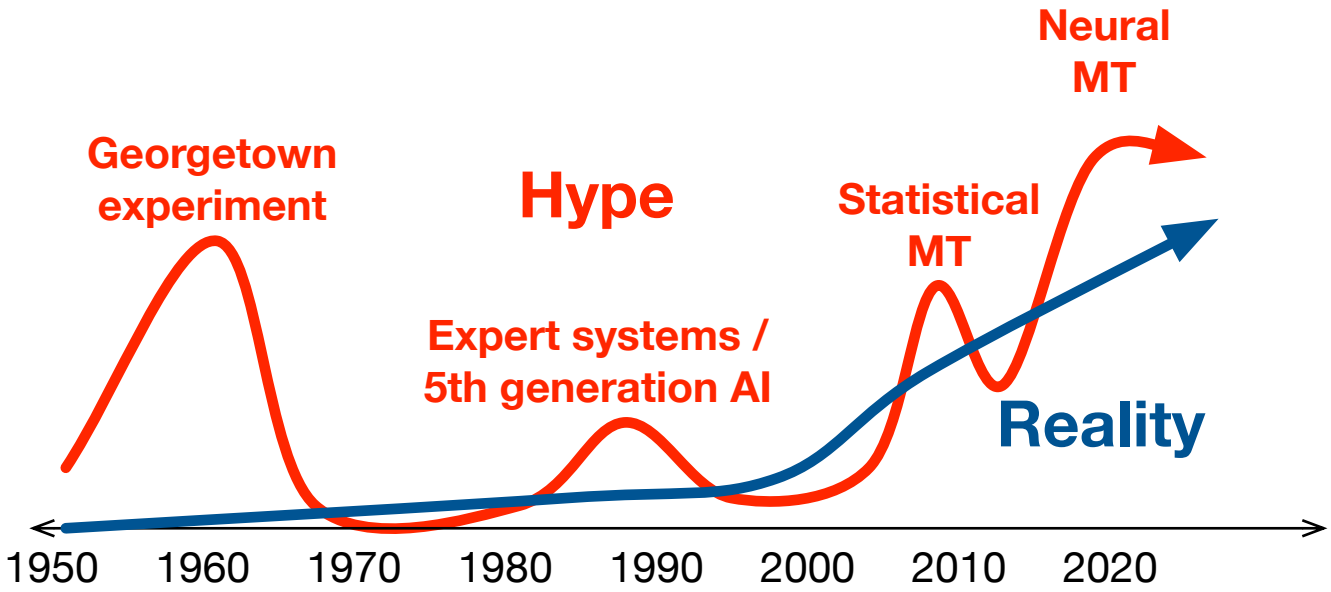
- Openly available models (NLLB, Llama, Qwen, ...)

- Open evaluation campaign: promote honest detailed system reports

- Long standing ideal: a graduate student should be able to participate
  — great ideas come from a diverse pool of researchers

Be Humble

- Long-standing awareness of hype outrunning reality



- Focus on “good enough”, enabling abilities
- Promote sharing and progress, not hype

Thank You

questions?