Summarize Large Text using NLP on NVIDIA P3 Instance

GTC | S9628 | Mar, 2019


"Summarization is the jungle of NLP"

Kristof Schum

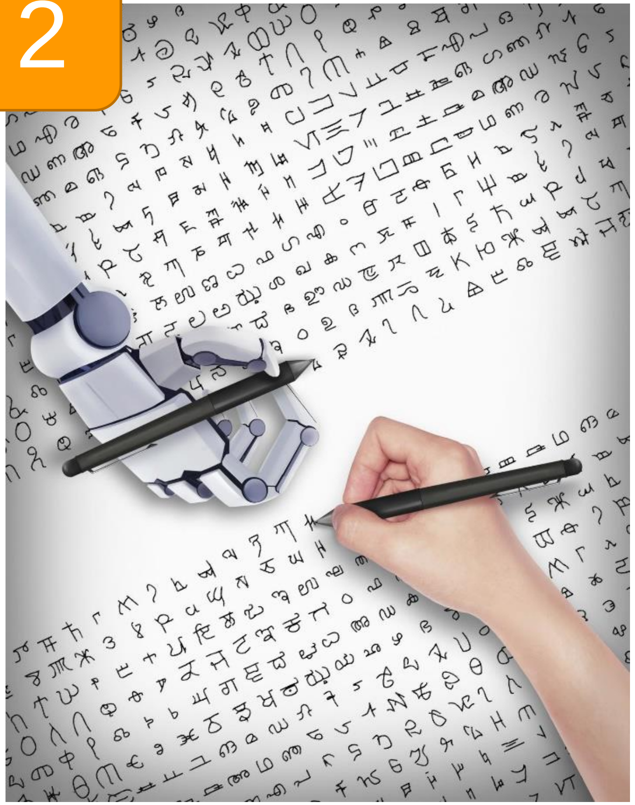
Global Segment Leader
Machine Learning
AWS Partner Network

- From consulting to ML PM
- Automated Insights
- Summarization from Wharton
- Teach Summarization at MLU

1

GPU up

2



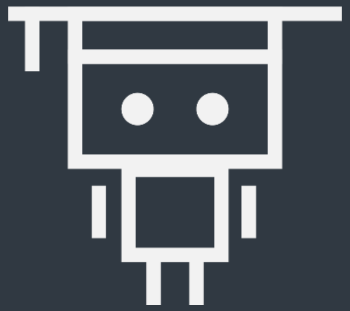
Teach

3

Innovate

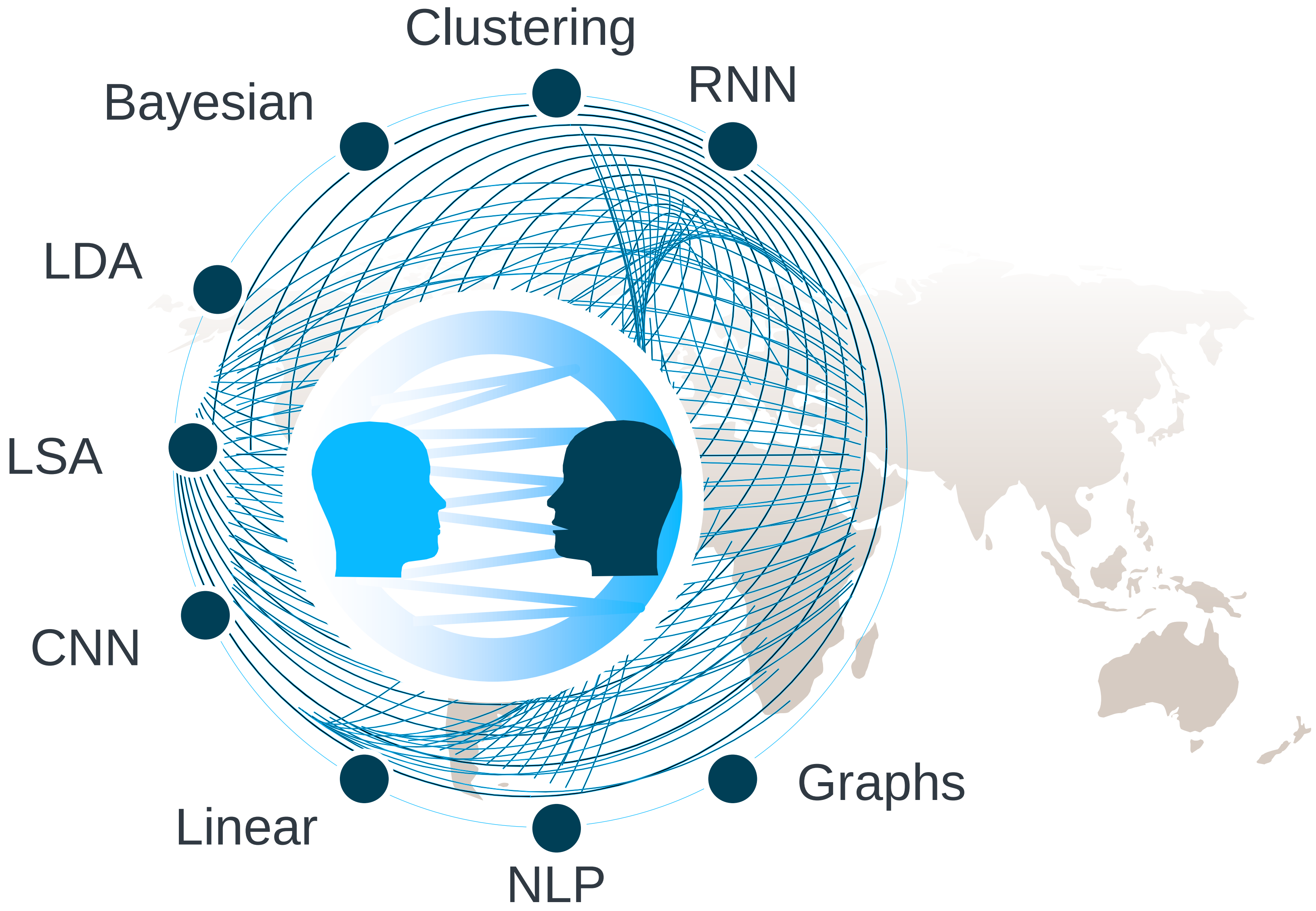
Why bother?

Summarization is not as fundamental and immediately applicable as a feed-forwarded neural net or XGBoost.



1. Trending

2. Multifaceted



3. Easy to innovate

Imagine you did not have time to take notes

+

=

Amazon Transcribe

Amazon Sagemaker

Notes Instantly

Agenda for today

Evolution

Of Automatic Text Summarization

Paraphrase

Methods in the field of
Natural language generation
That provide new text as summary

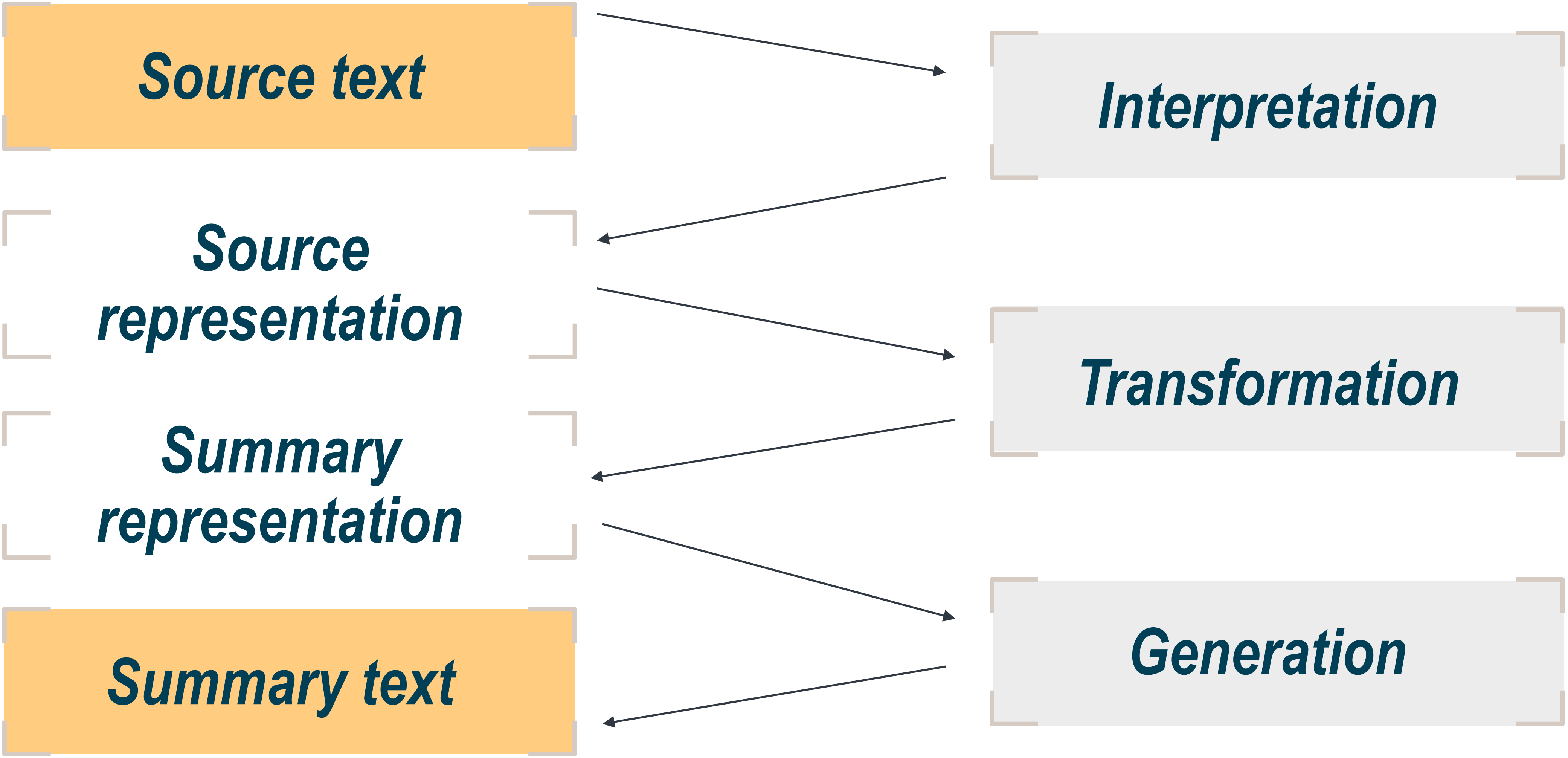
Statistical

Methods that are focused on finding and extracting the
most expressive as-is sentences in the text

1. Evolution

“A **reductive transformation** of source text to summary text through **content condensation** by selection and/or generalization on what is **important in the source**.”

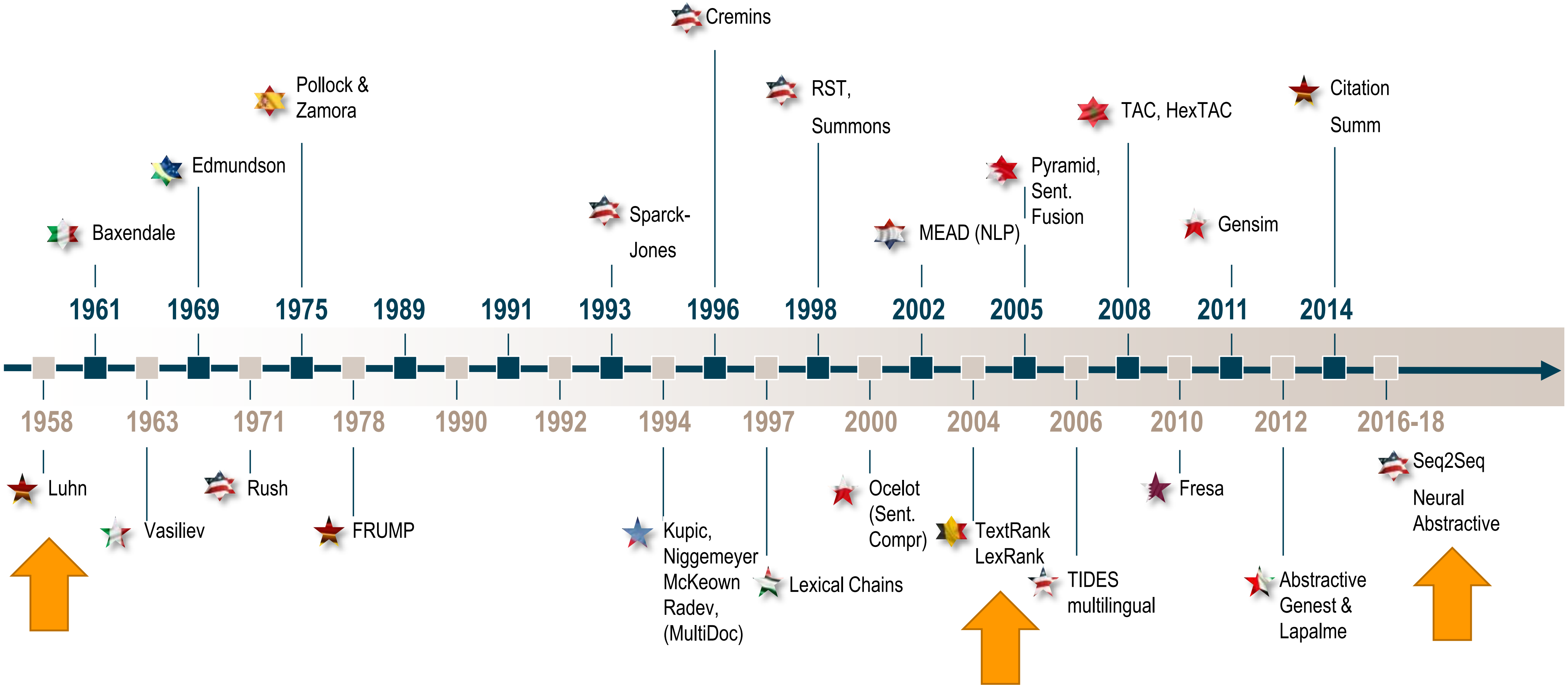
Schematic summary processing model

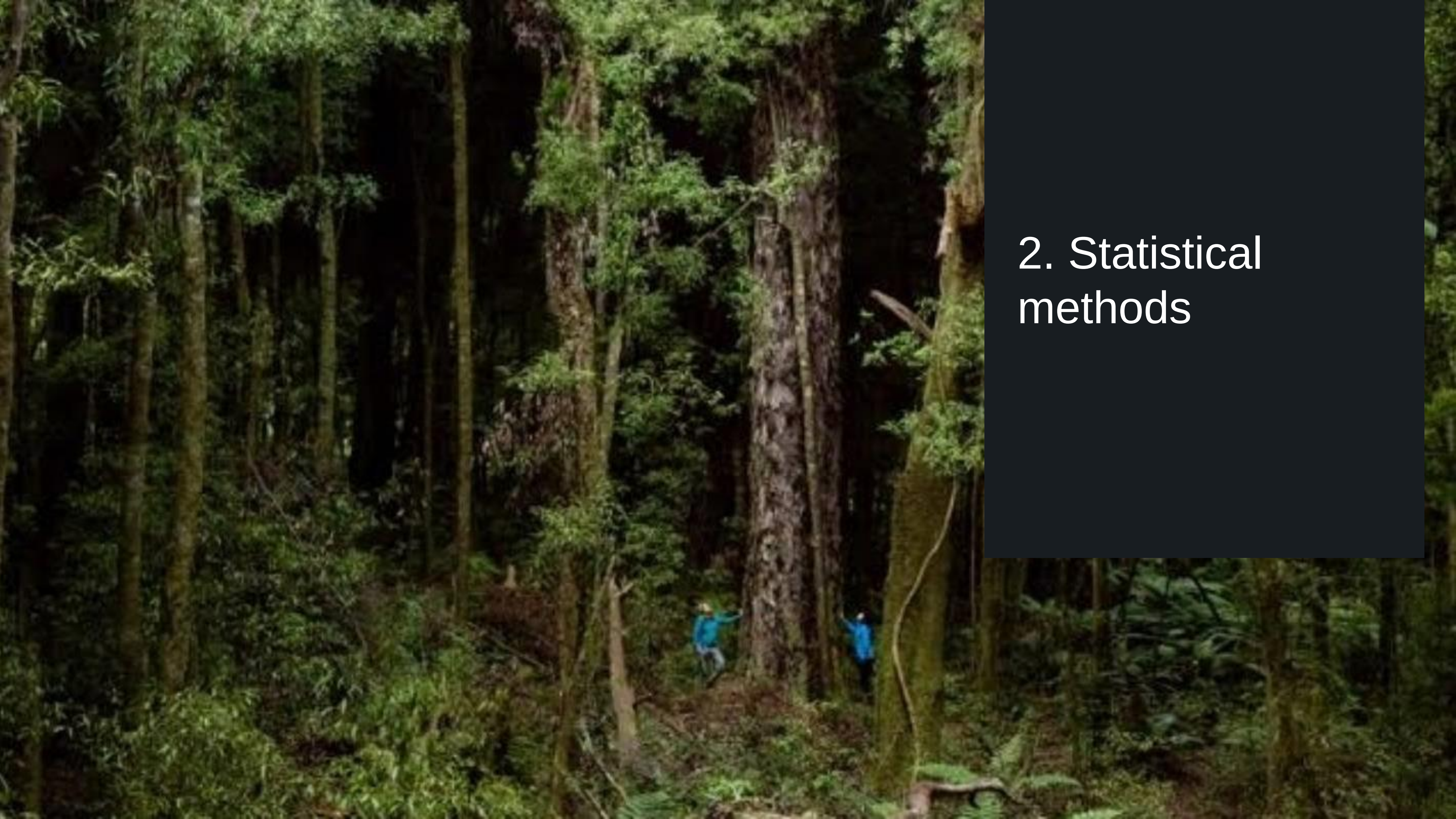


‘Genres’ of Summary?

- Indicative *vs*. informative
  ...used for quick categorization vs. content processing.
- Extract *vs*. abstract
  ...lists fragments of text vs. re-phrases content coherently.
- Generic *vs*. query-oriented
  ...provides author’s view vs. reflects user’s interest.
- Background *vs*. just-the-news
  ...assumes reader’s prior knowledge is poor vs. up-to-date.
- Single-document *vs*. multi-document source
  ...based on one text vs. fuses together many texts.

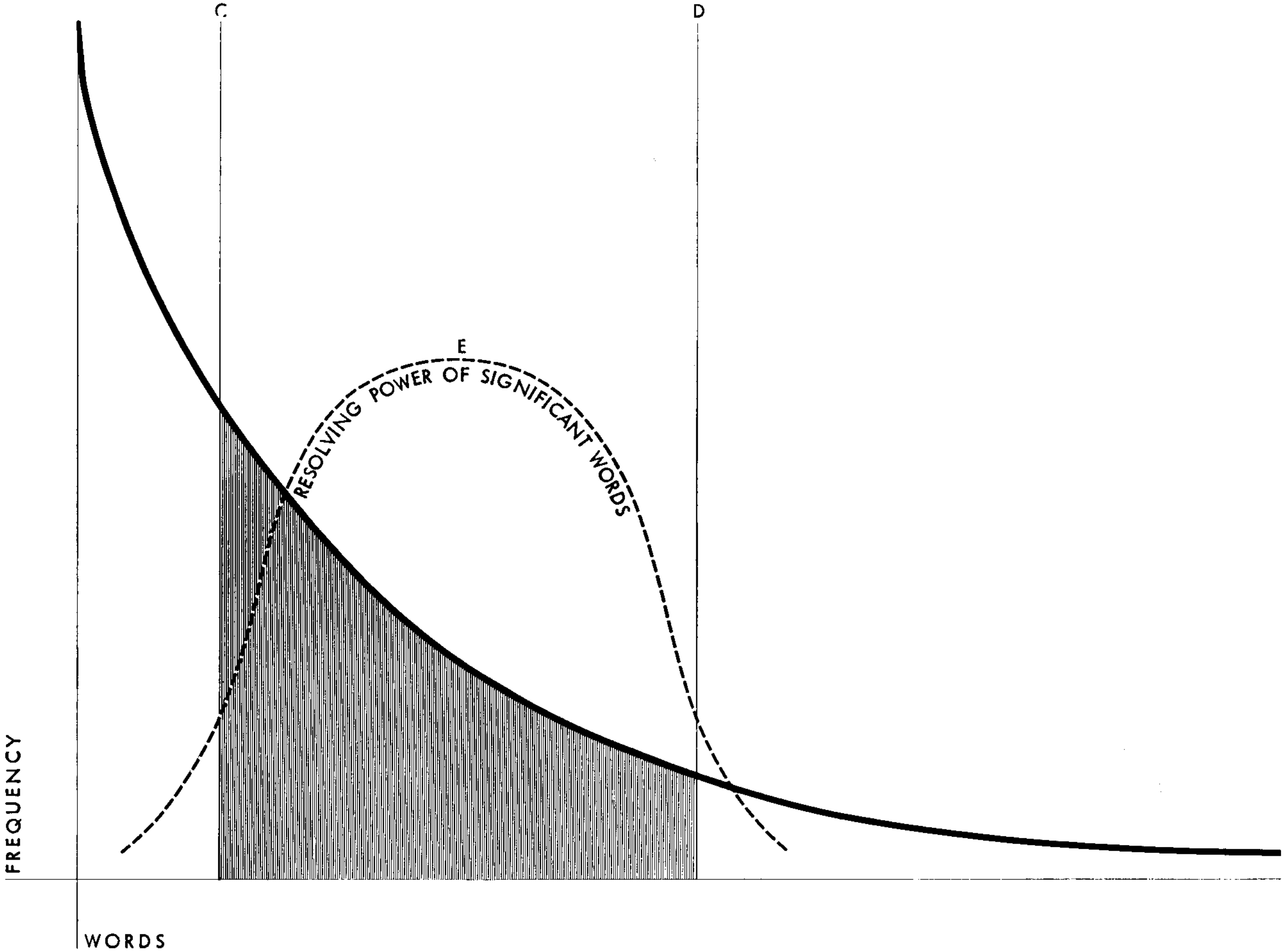
Evolution of methods





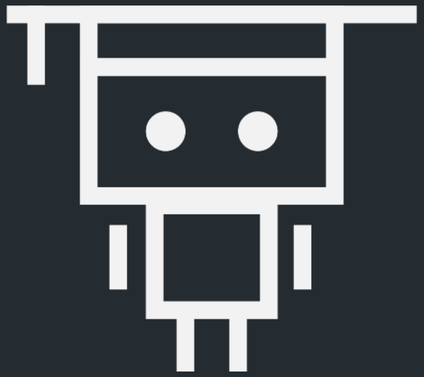
2. Statistical methods

The father of information retrieval



Let’s give it an easy time

Demo



5 sentences generated from the article:

* It’s that time of year again.

* This conference always hosts a smorgasbord of informative keynotes, exhibitors, and hands-on sessions, on a wide variety of topics.

* **The program will include a women-led panel session, women-only DLI sessions, and a networking reception.**

* The conference will also focus on up-and-coming fields such as finance, healthcare, and telco.

* The conference continues to expand, with more sessions, more exhibitors, and more emergent topics of discussion (healthcare, telco, finance, etc.


Feb 7, 2019, 10:30am

NVIDIA Gears Up For An Even Larger GTC 2019

Patrick Moorhead Contributor
Enterprise & Cloud
I write about disruptive companies, technologies and usage models.

GTC 2018 attendees NVIDIA

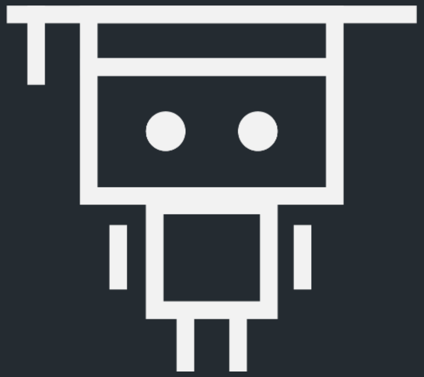
It's that time of year again. Every spring NVIDIA kicks off its annual series of GPU Technology Conferences (GTC) with a real "humdinger" of an event held in San Jose. Last year, I wrote that GTC 2018 was the place to be if you are in any way involved in AI or

(link)

amazon MACHINE LEARNING UNIVERSITY

Let’s give it a hard time

Demo



An excerpt from *The Blah Story, Volume 15*:

“Her blah didn’t blah blah to blah some blah advantages. The blah was blah and blah blah, but she blah quite a blah blah blah. Nevertheless, the blah blah that blah gave the blah blah was blah of blah, irony, and blah blah. When blah had blah blah that blah was likely to blah blah a blah once blah she blah no blah of her blah. She blah to blah old blah blah more blah than blah.”

The Blah Story

11.3M words

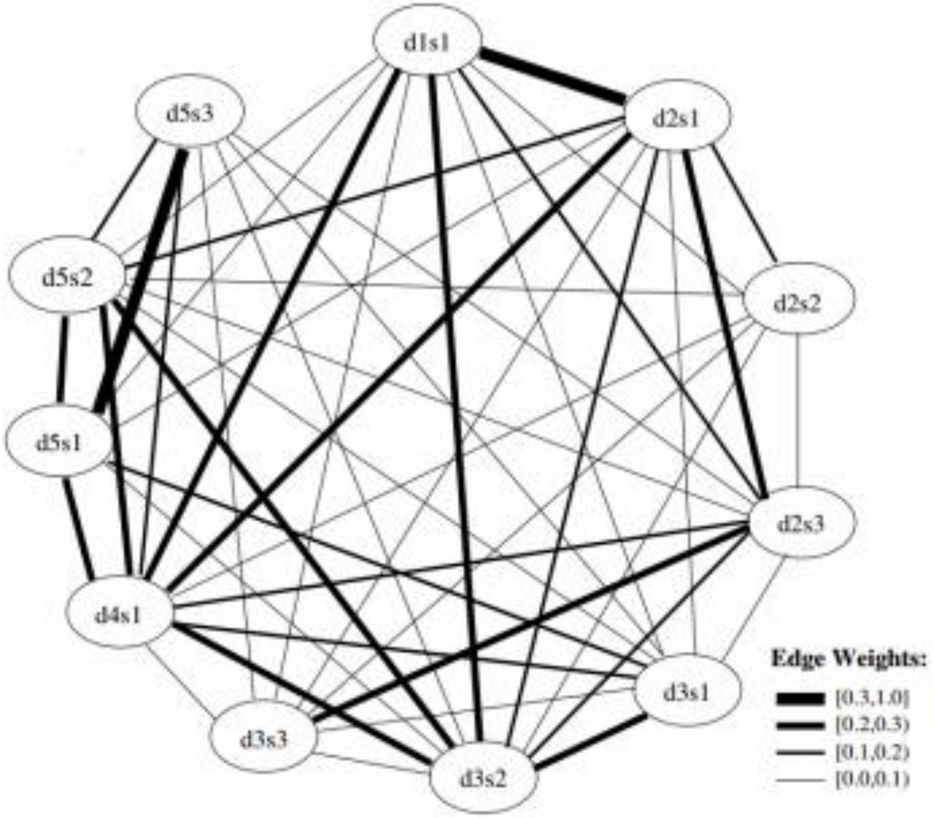
17,868 pages

Sentences generated from the Lord of The Rings:

* He looked at the great walls, and the towers and brave banners, and the sun in the high sky, and then at the gathering gloom in the East; and he thought of the long fingers of that Shadow: of the ores in the woods and the mountains, the treason of Isengard, the birds of evil eye, and the Black Riders even in the lanes of the Shire - and of the winged terror, the Nazgyl.

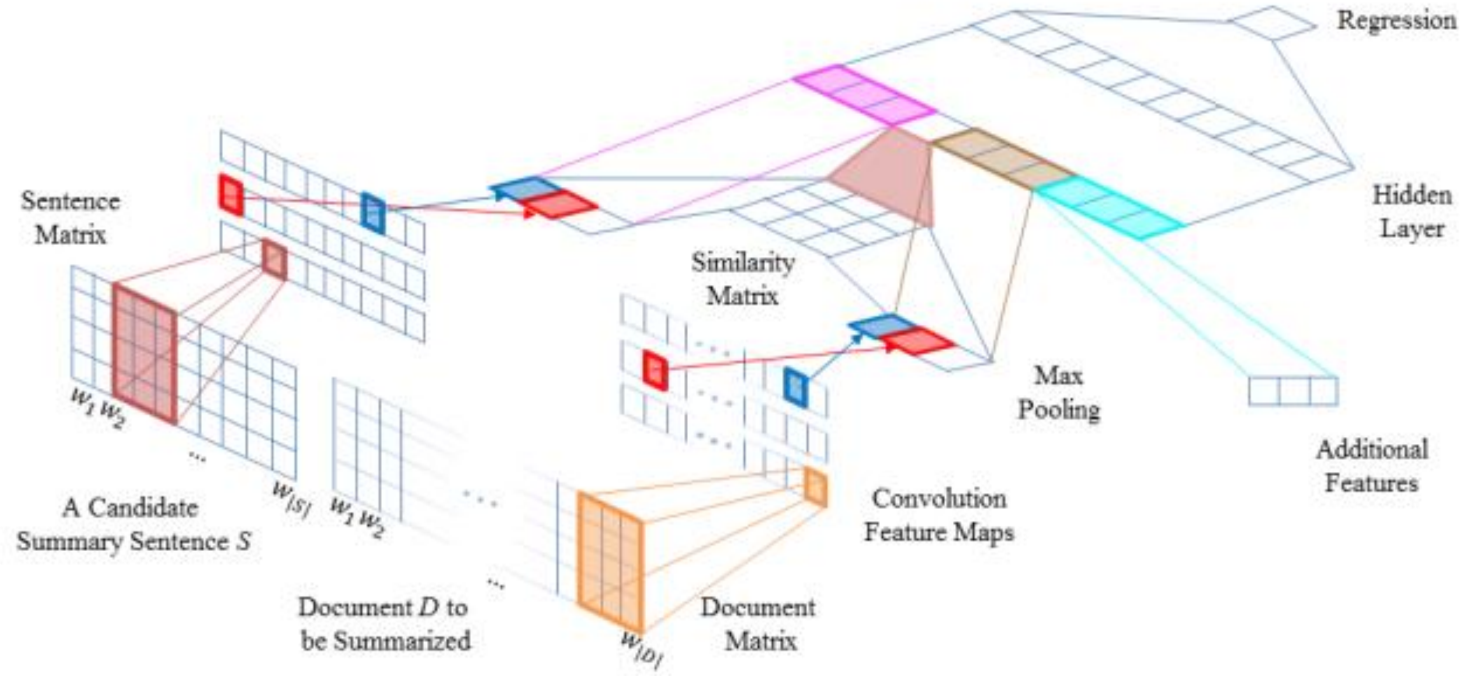
… [4 more]

A more sophisticated statistical method



Source: Rada – Tarau 2014

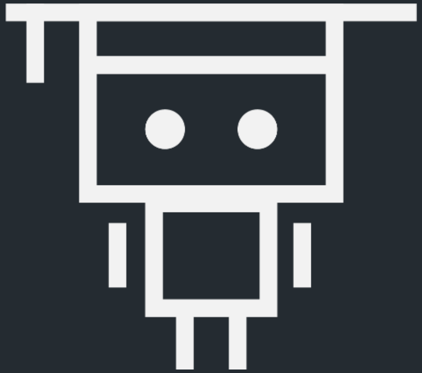
An alternative: similarity with CNNs



Source: Zhang, Er, Pratama - 2016

Let’s give TextRank an easy time

Demo



5 sentences generated from the article:

- The story holds true for this year’s event (held March 17-21), with NVIDIA promising to shine a spotlight on all the impactful applications of AI, including robotics and autonomous vehicles with a larger keynote area and more exhibitors.

- This year’s conference speaker roster features a who’s who in AI and deep learning, with experts from industry leaders such as Amazon, Alibaba, Google, NASA, Oak Ridge National Labs, IBM, Verizon, Volvo, PayPal, and many, many more.

- NVIDIA’s tech rock star CEO Jensen Huang will be delivering his keynote (no doubt in his signature leather jacket) on Monday afternoon, at the San Jose State event center, which seats 5,000 (2,000 more than last year’s venue).

- NVIDIA says 9 of the world’s top 12 telco companies will be attending and presenting at this year’s GTC, as well as 4 of the top 5 medical research universities and 5 of the top 7 radiology departments.

- NVIDIA promises more Deep Learning Institute (DLI) coverage this year, with six all-day workshops (including developer certification), and over 100 DLI sessions all said and told.


Feb 7, 2019, 10:30am

NVIDIA Gears Up For An Even Larger GTC 2019

Patrick Moorhead Contributor

Enterprise & Cloud

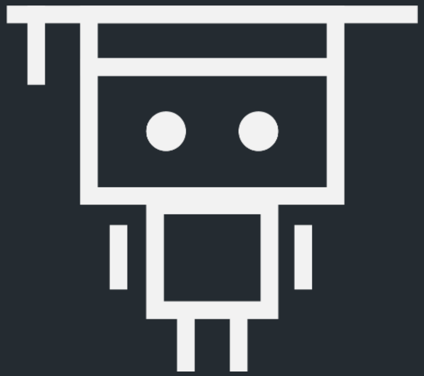
I write about disruptive companies, technologies and usage models.

It's that time of year again. Every spring NVIDIA kicks off its annual series of GPU Technology Conferences (GTC) with a real "humdinger" of an event held in San Jose. Last year, I wrote that GTC 2018 was the place to be if you are in any way involved in AI or

(link)

Let’s give Textract a LOTR time

Demo

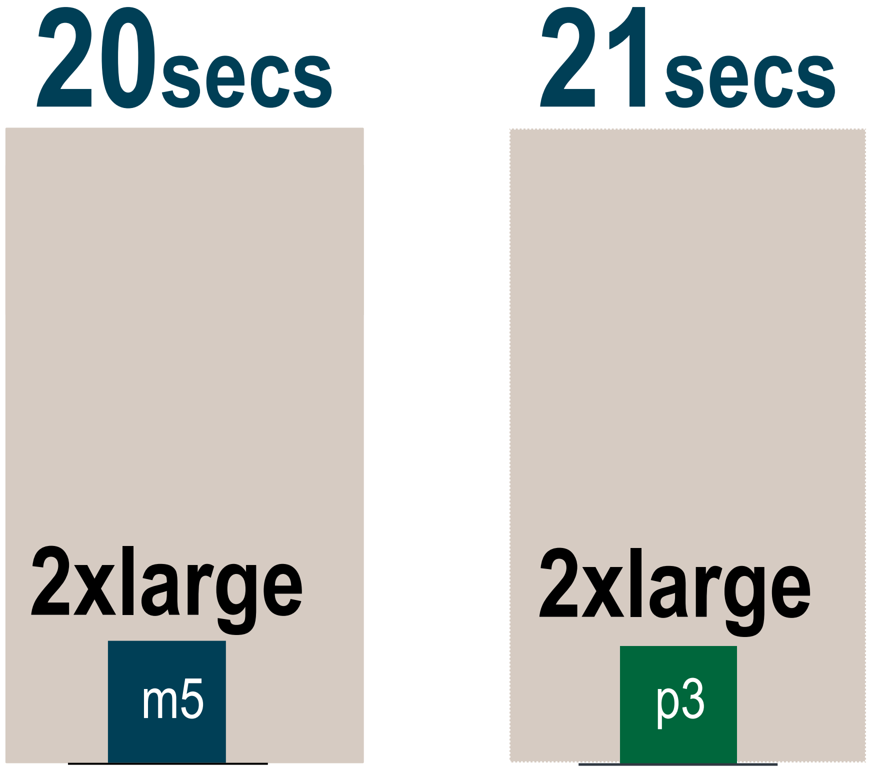


Sentences generated from the Lord of The Rings:

The Hobbits named it the Shire, as the region of the authority of their Thain, and a district of well-ordered business; and there in that pleasant comer of the world they plied their well-ordered business of living, and they heeded less and less the world outside where dark things moved, until they came to think that peace and plenty were the rule in Middle-earth and the right of all sensible folk.

… 4 more

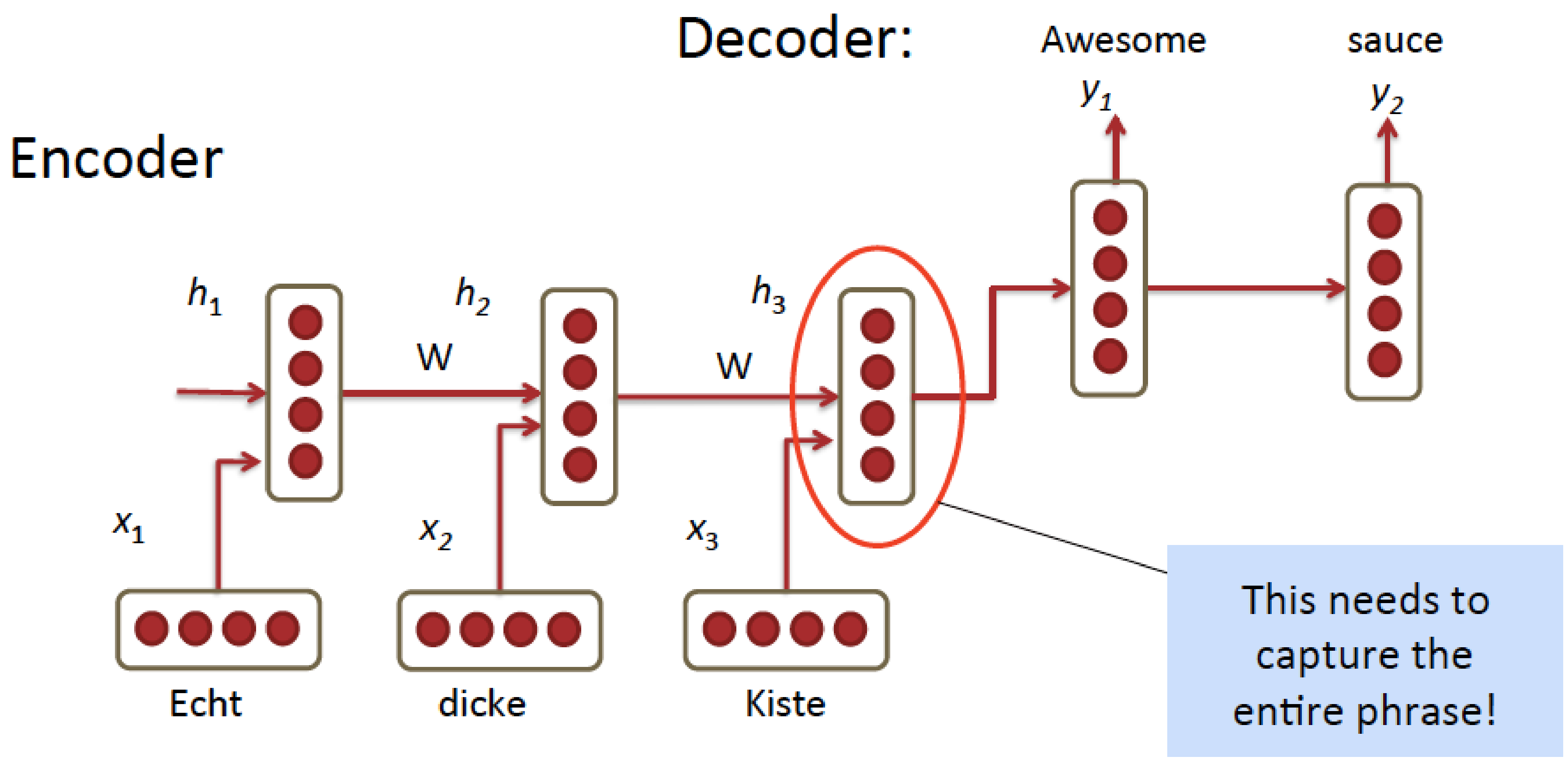
No deep learning, no need for P3



amazon MACHINE LEARNING UNIVERSITY

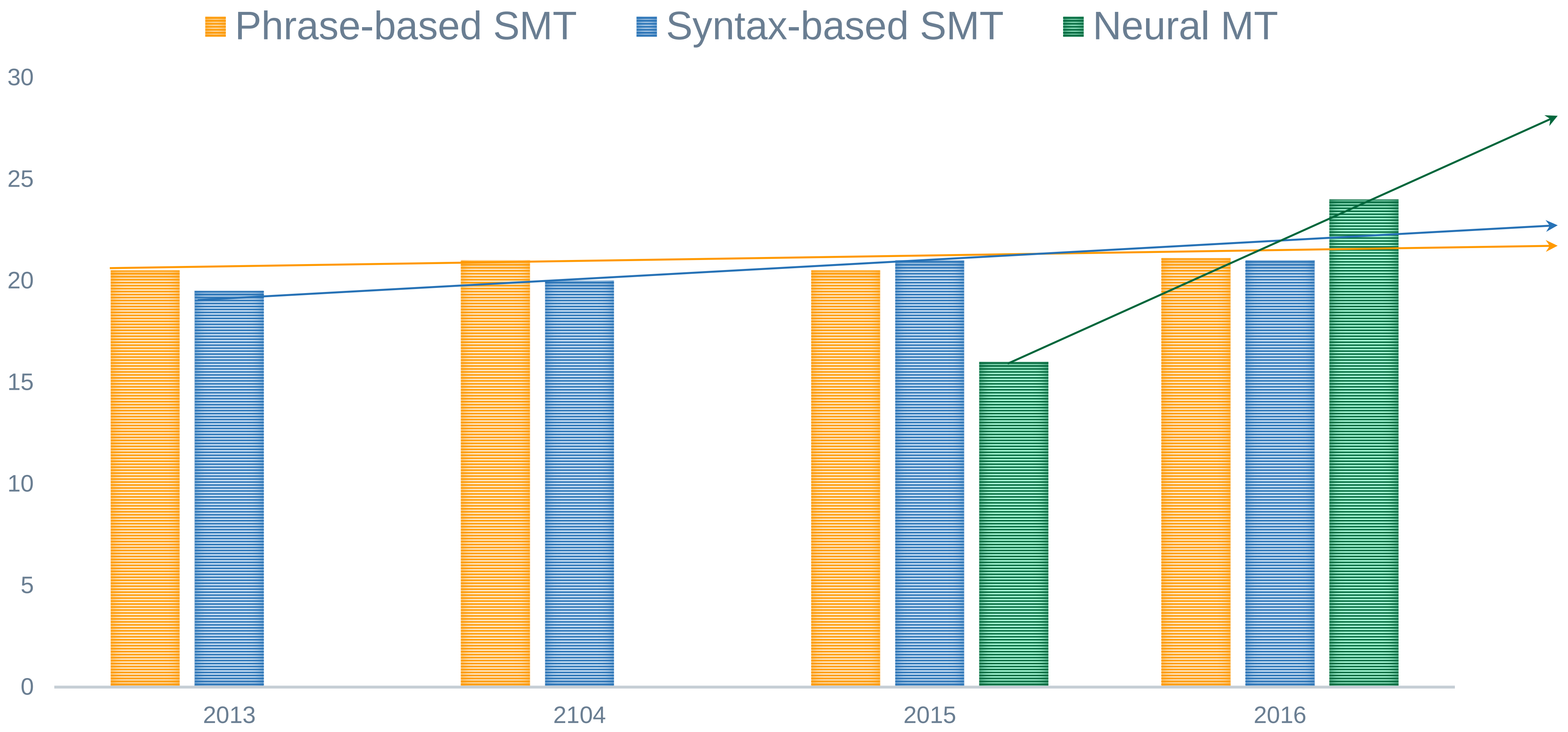
3. Paraphrasing method

Deep learning to the rescue - RNNs



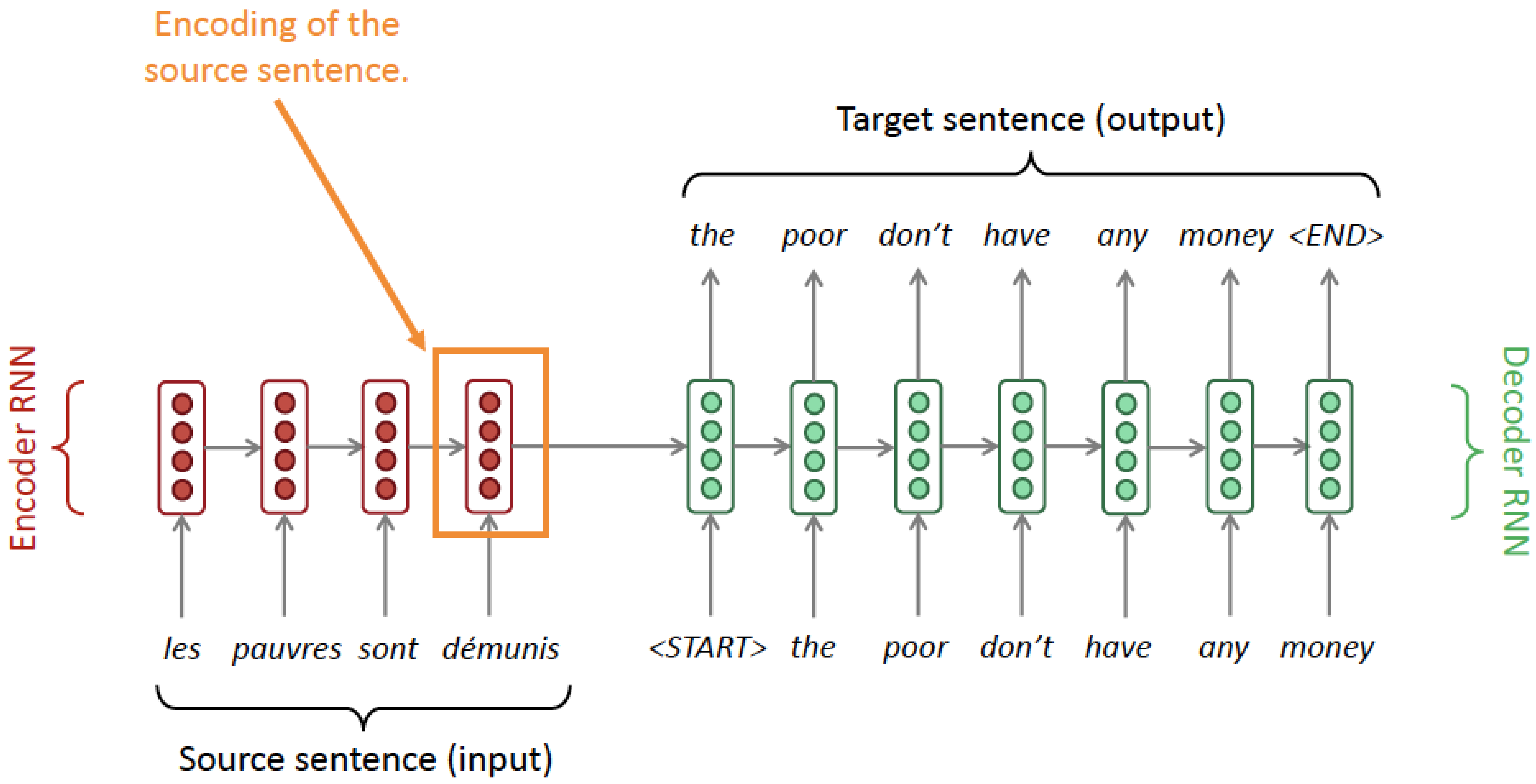
MT progress over time

[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal]

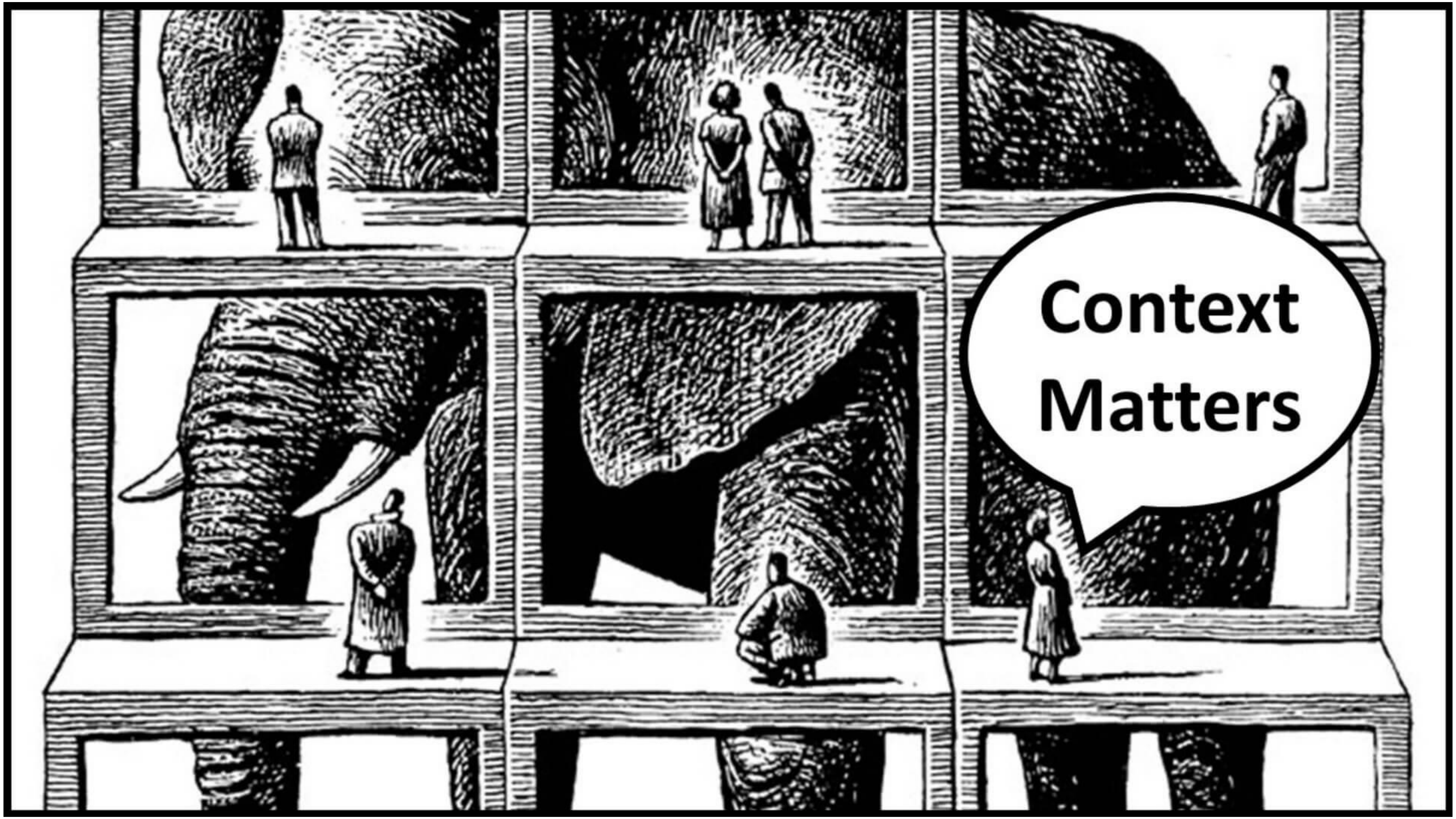


Source: Meta Forum 2016 - Sennrich

Sequence-to-sequence: the bottleneck problem



Problems with this architecture?


Context
Matters

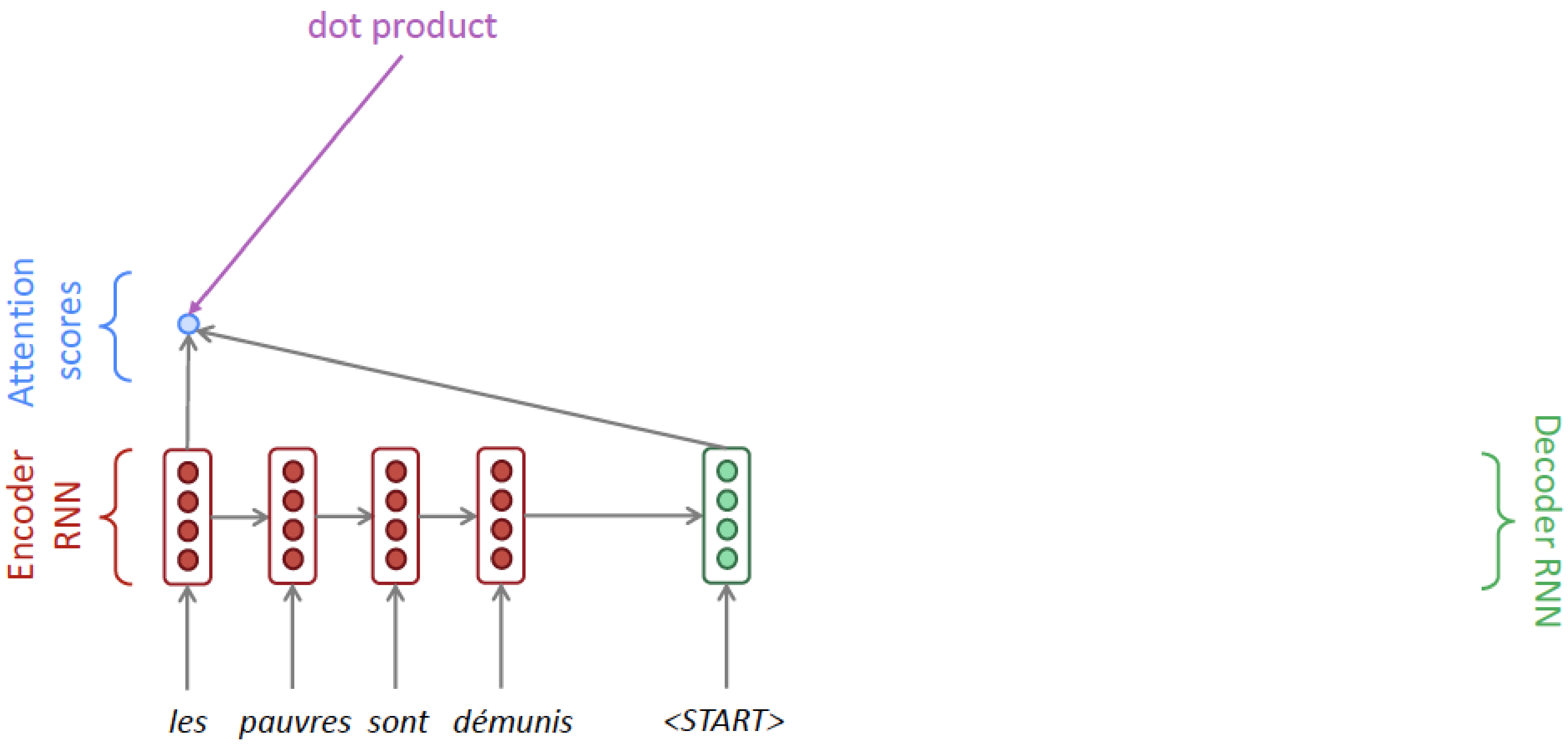
Attention is a *general* Deep Learning technique

More general definition of attention:

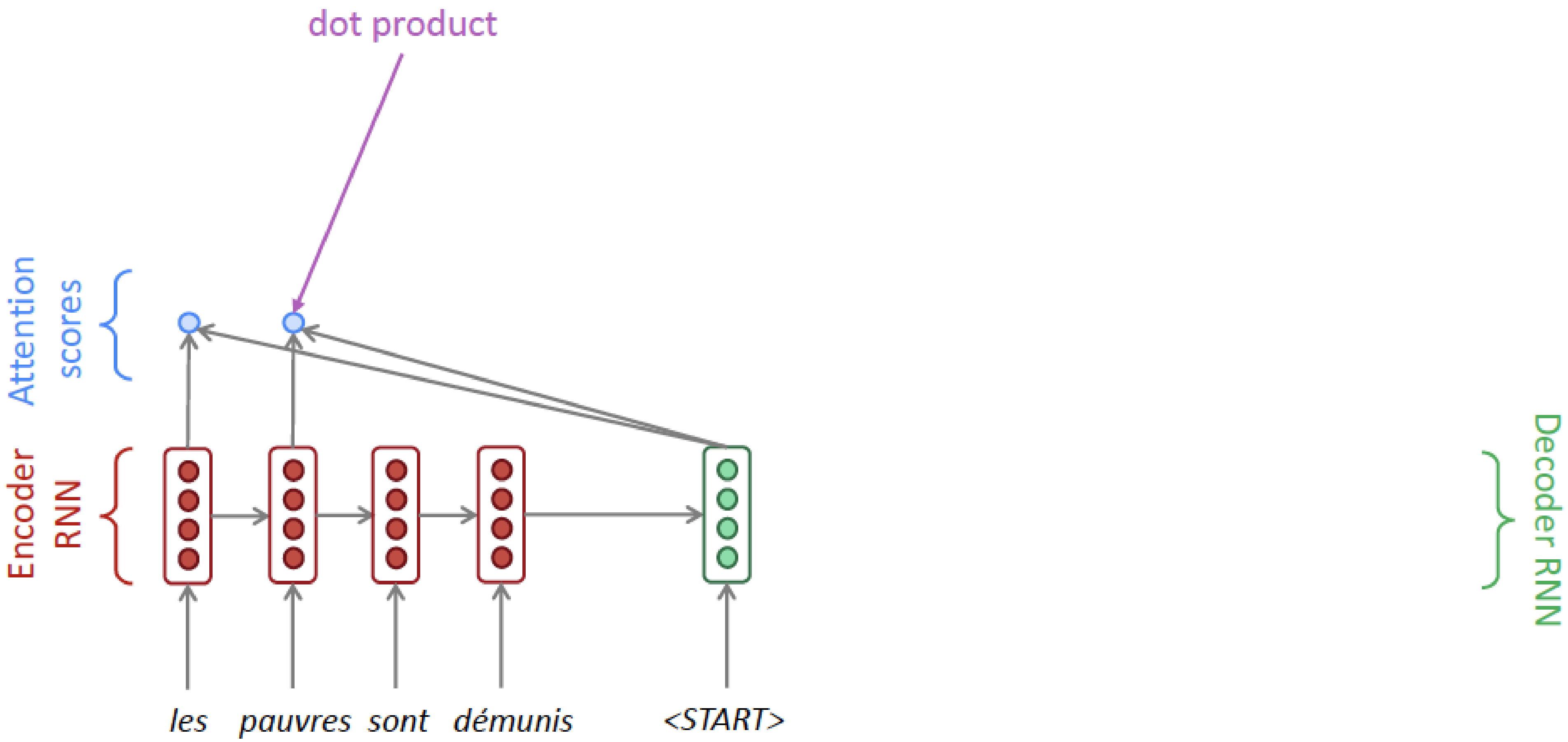
Given a set of vector *values*, and a vector *query*, **attention** is a technique to compute a weighted sum of the values, dependent on the query.

- **Intuition**:
  - The weighted sum is a *selective summary* of the information contained in the values, where the query determines which values to focus on.
  - Attention is a way to obtain a *fixed-size representation of an arbitrary set of representations* (the values), dependent on some other representation (the query).

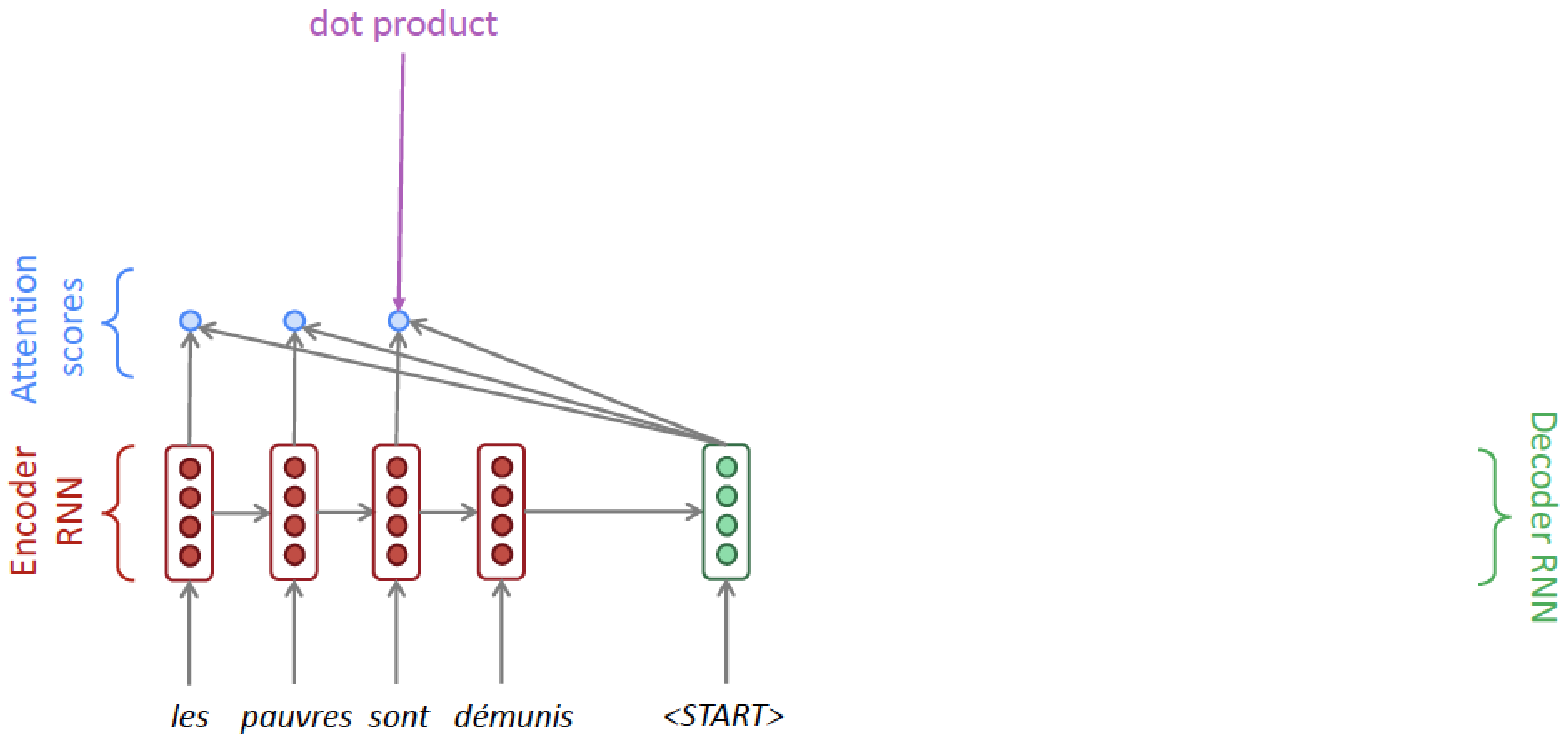
Sequence-to-sequence with attention



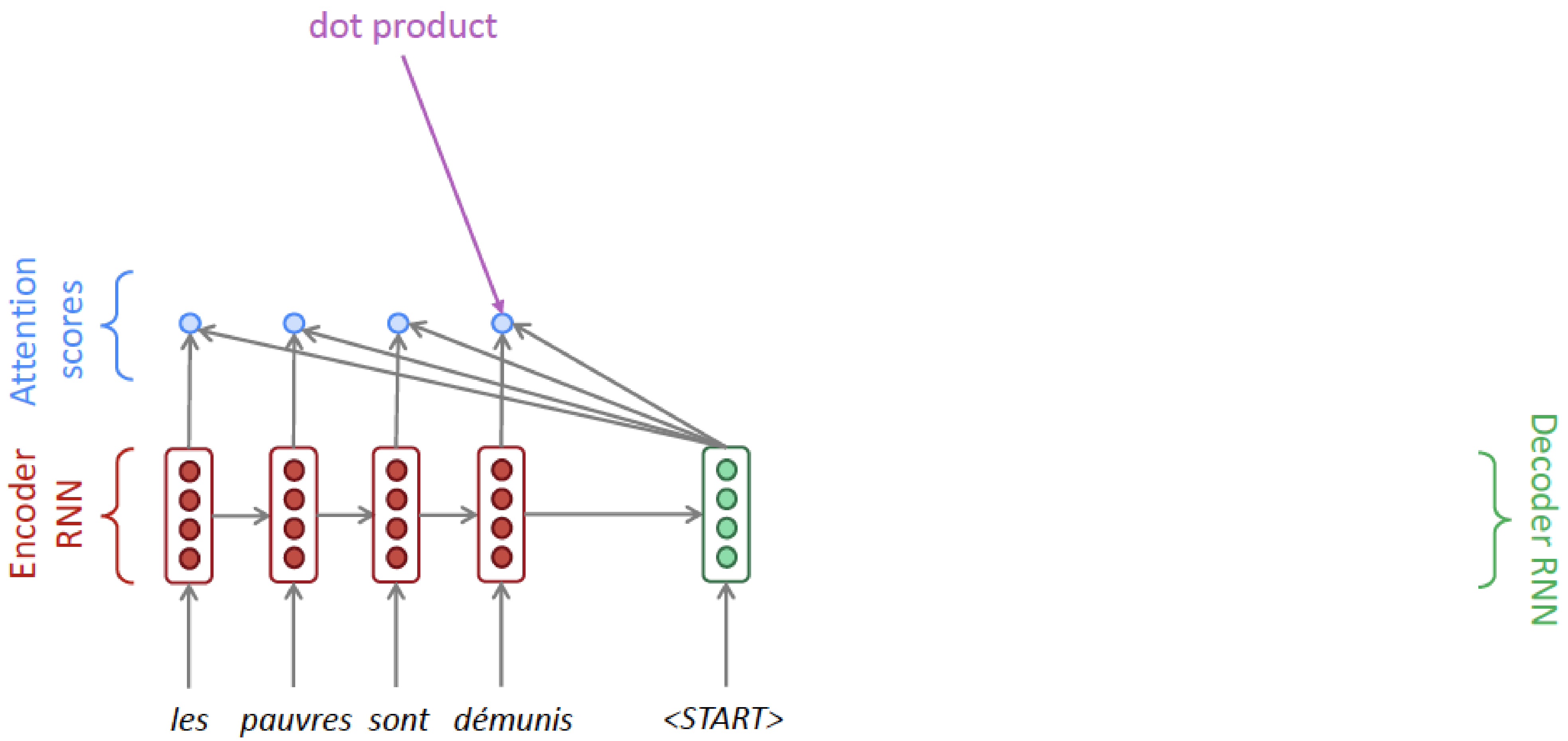
Sequence-to-sequence with attention



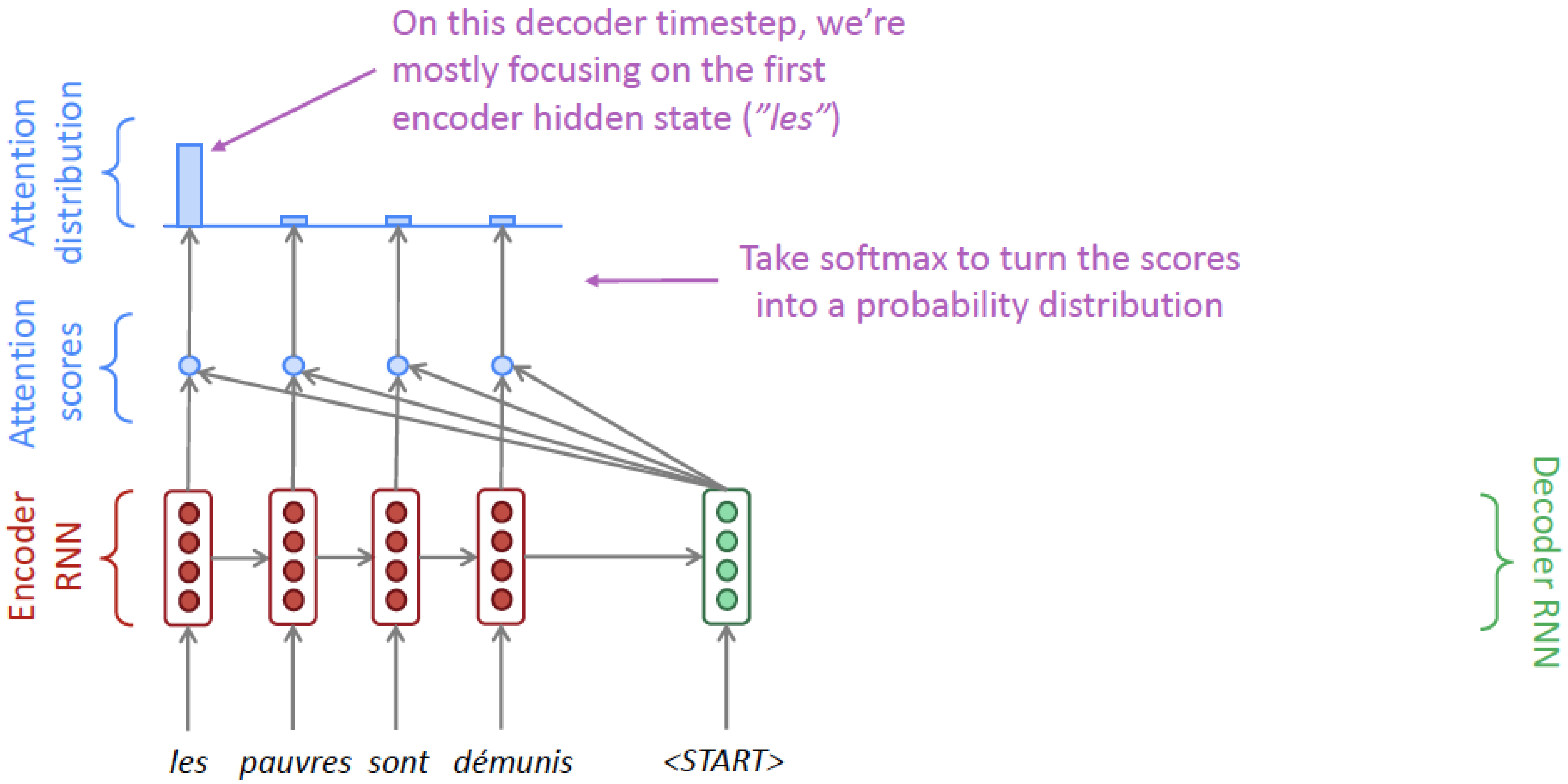
Sequence-to-sequence with attention



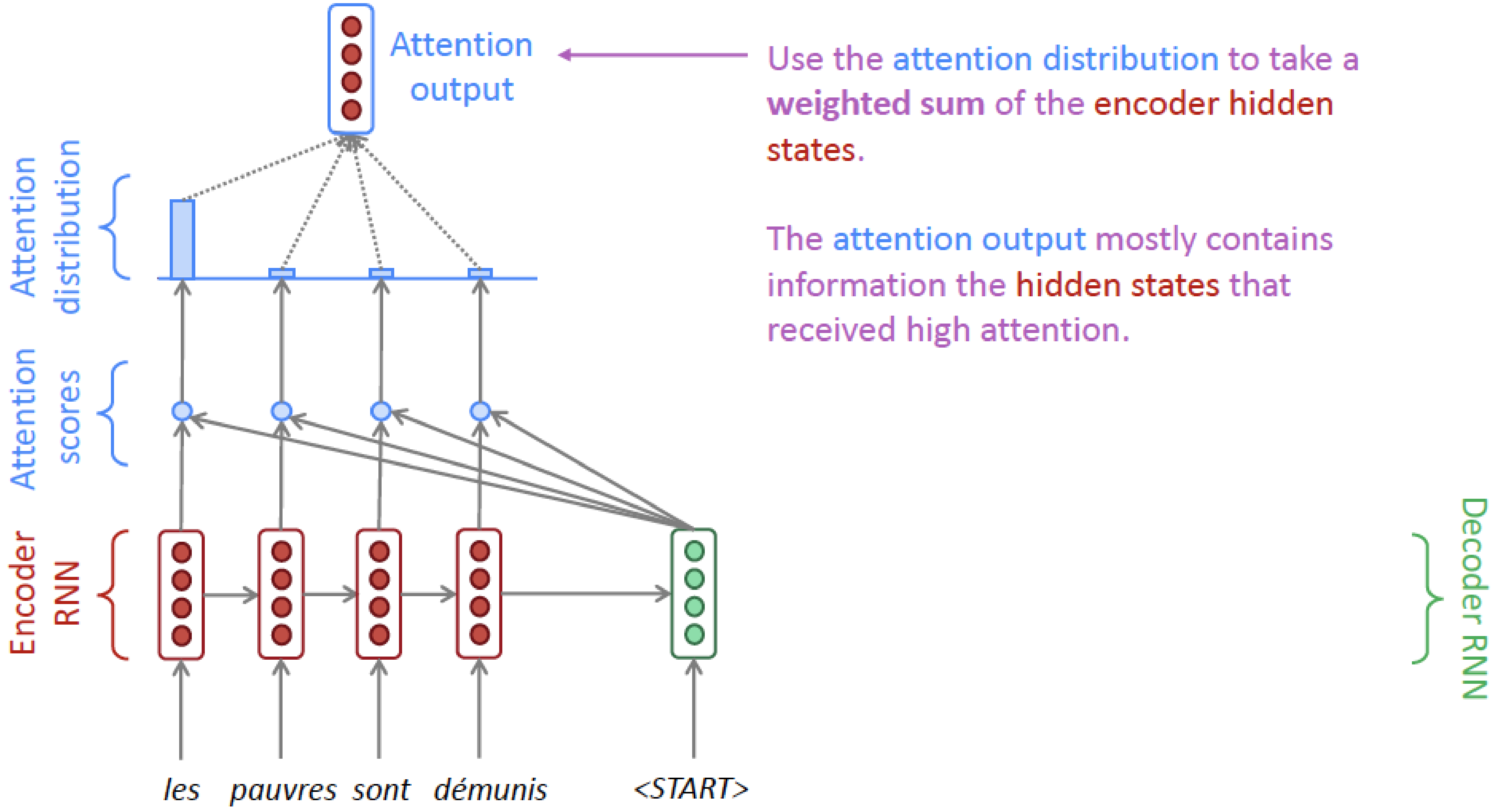
Sequence-to-sequence with attention



Sequence-to-sequence with attention



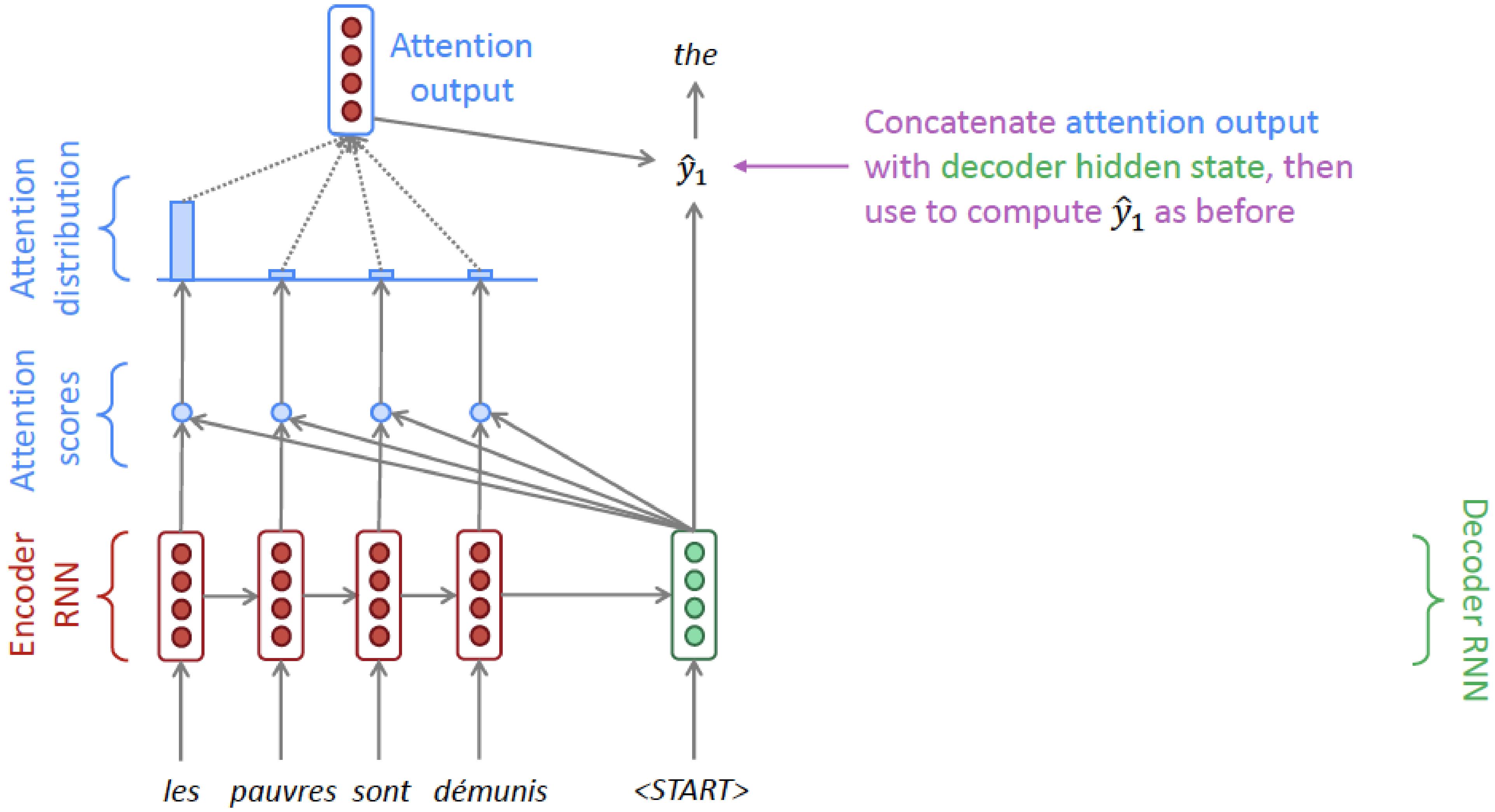
Sequence-to-sequence with attention



Use the attention distribution to take a **weighted sum** of the **encoder hidden states**.

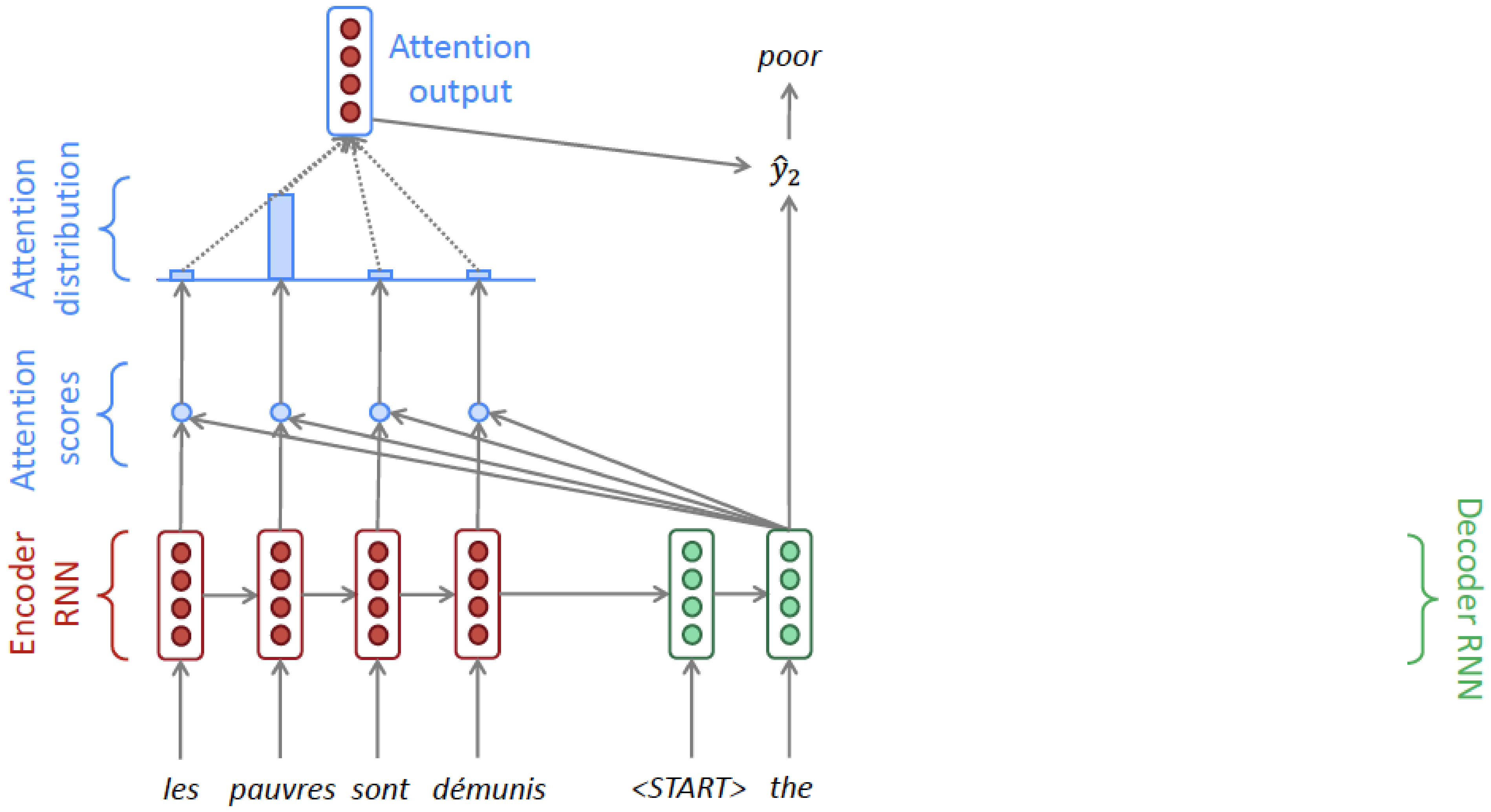
The attention output mostly contains information the **hidden states** that received high attention.

Sequence-to-sequence with attention

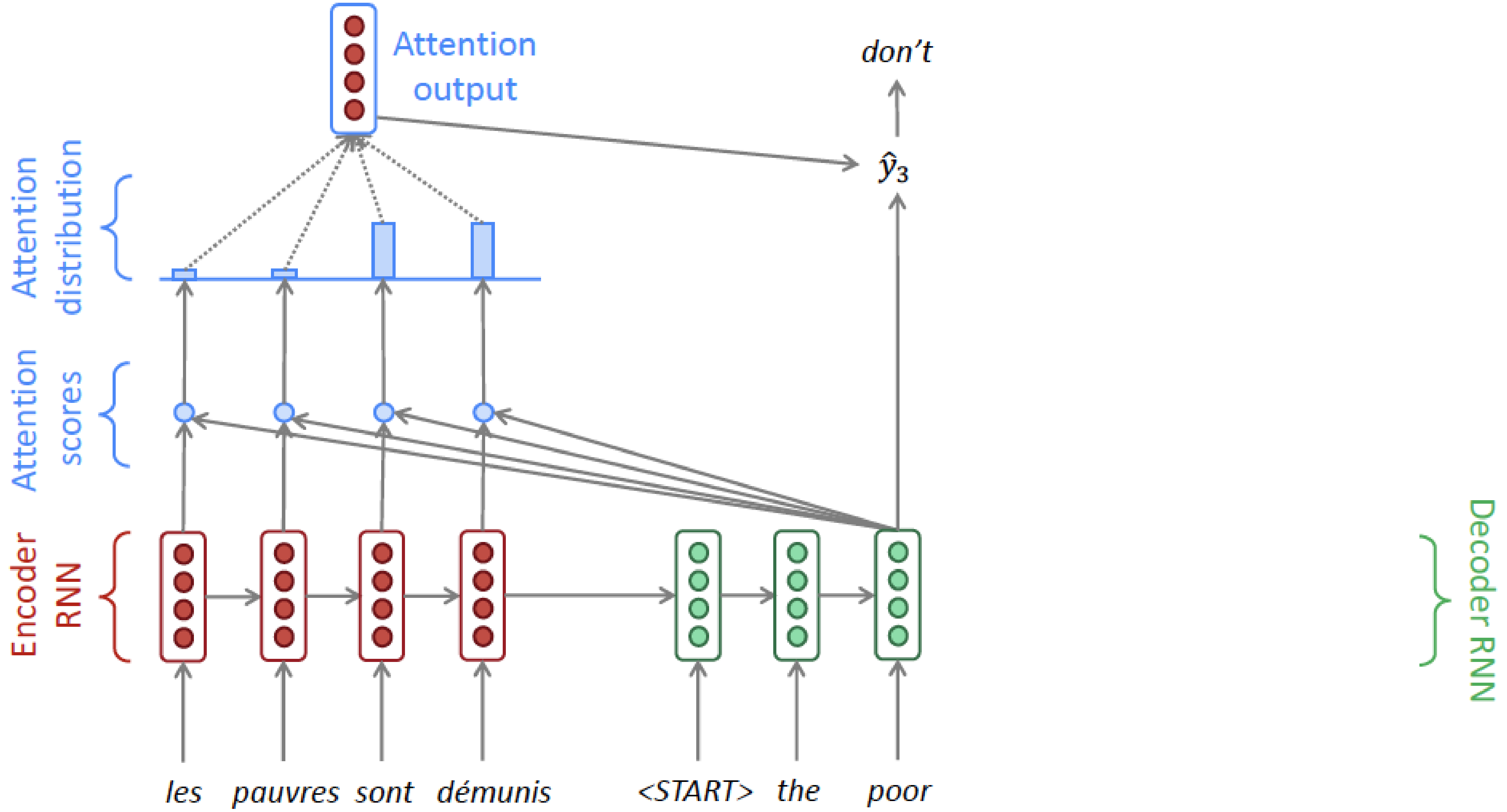


Concatenate attention output with decoder hidden state, then use to compute $\hat{y}_1$ as before

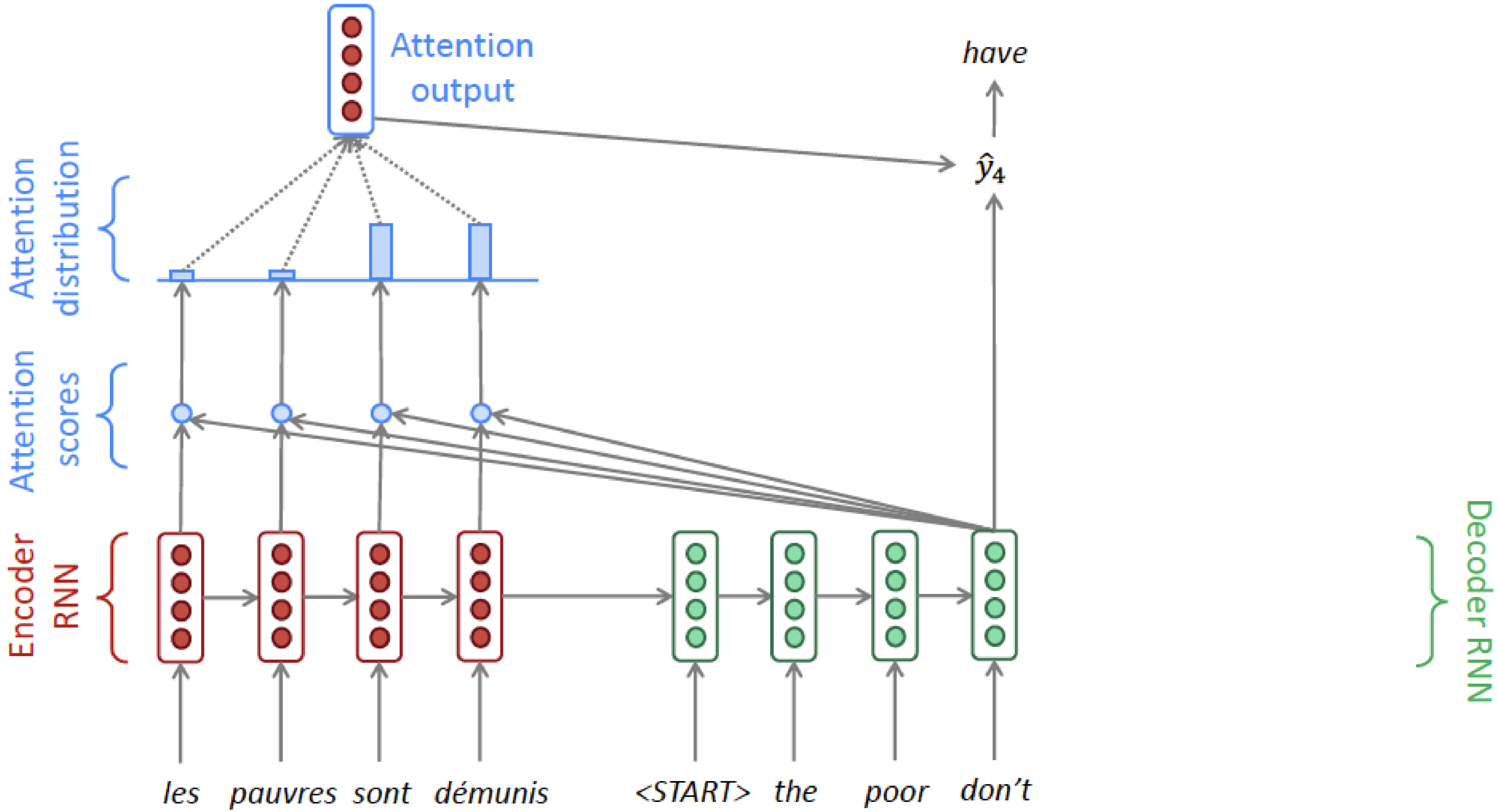
Sequence-to-sequence with attention



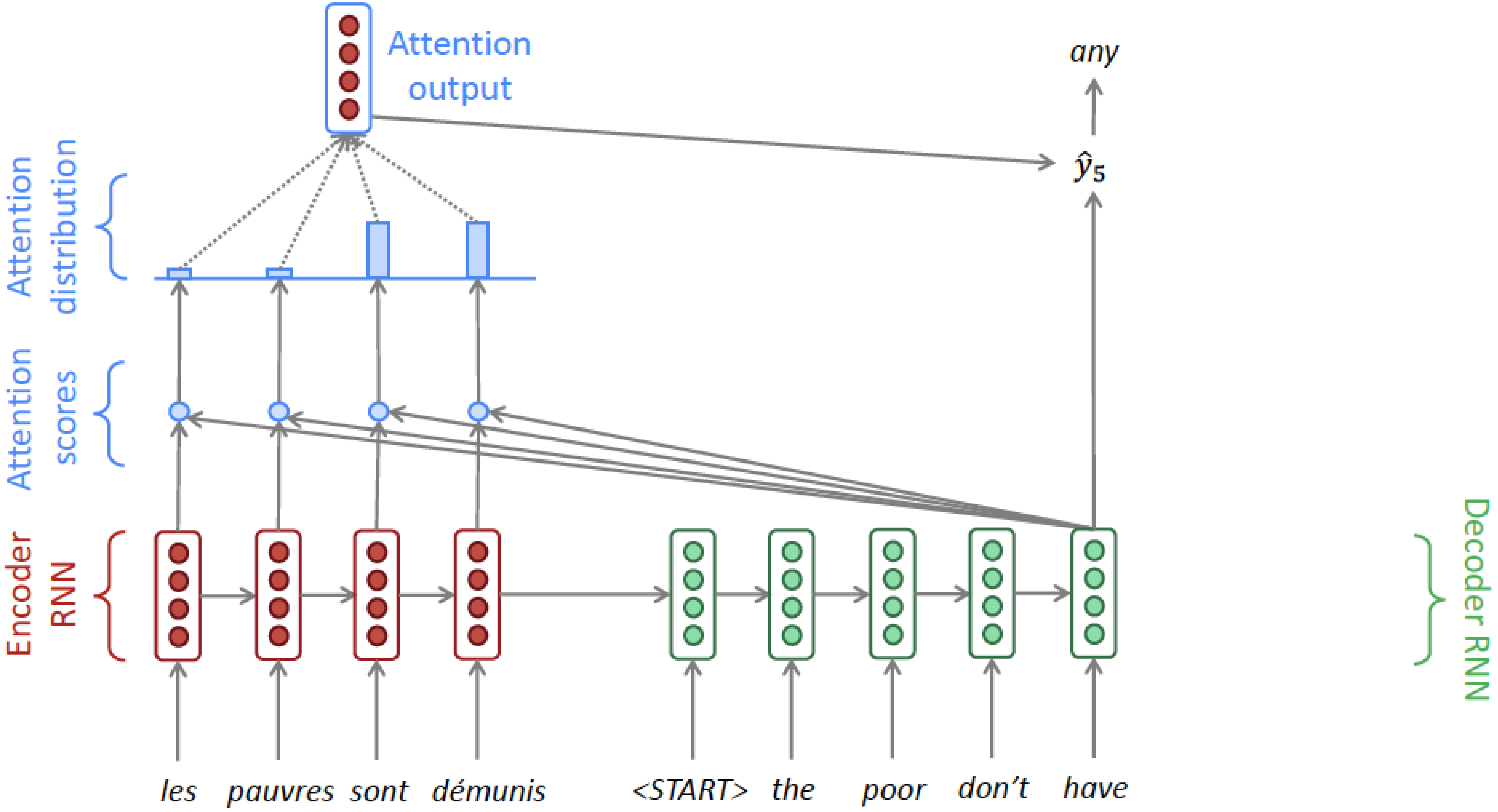
Sequence-to-sequence with attention



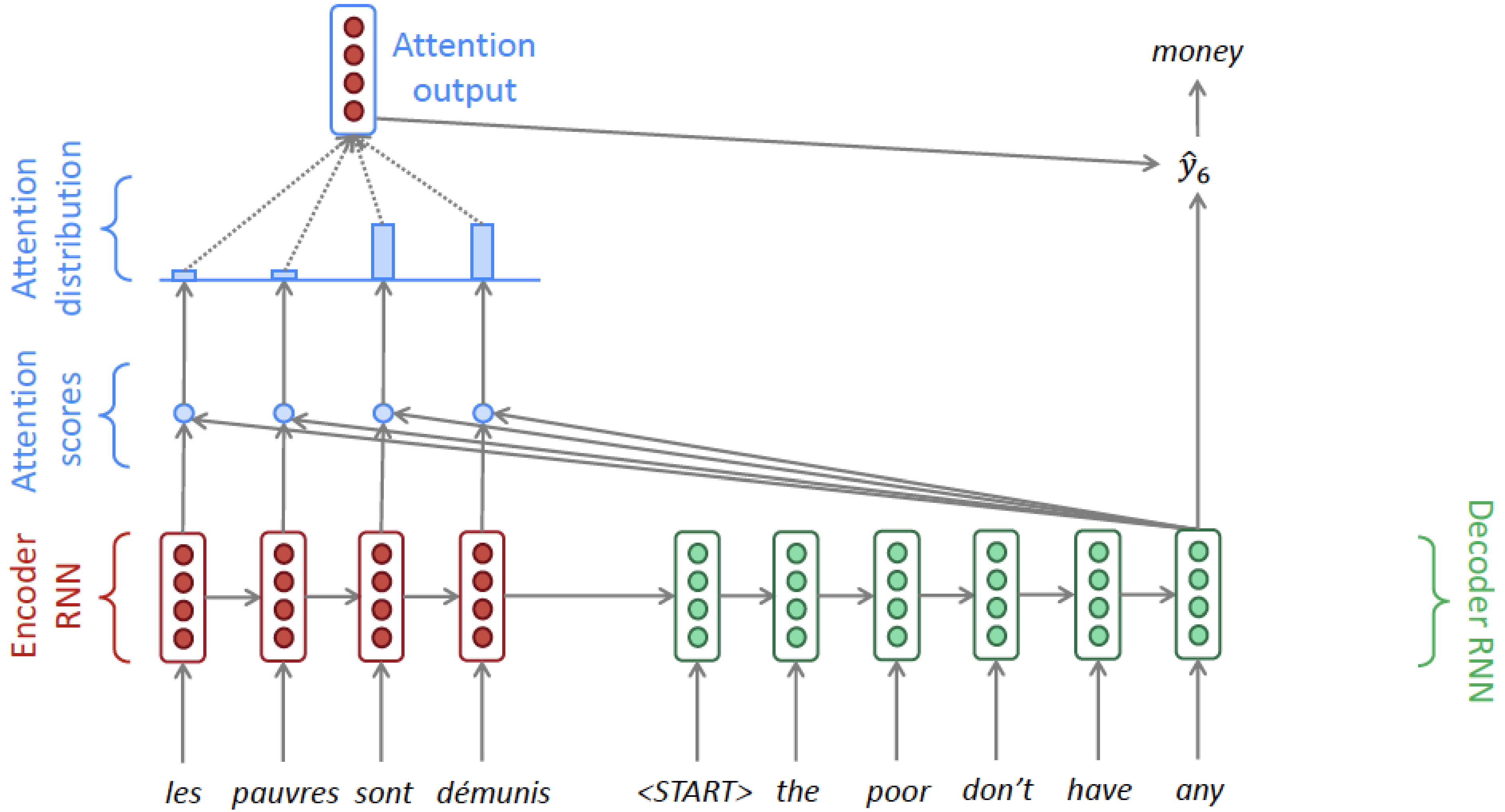
Sequence-to-sequence with attention



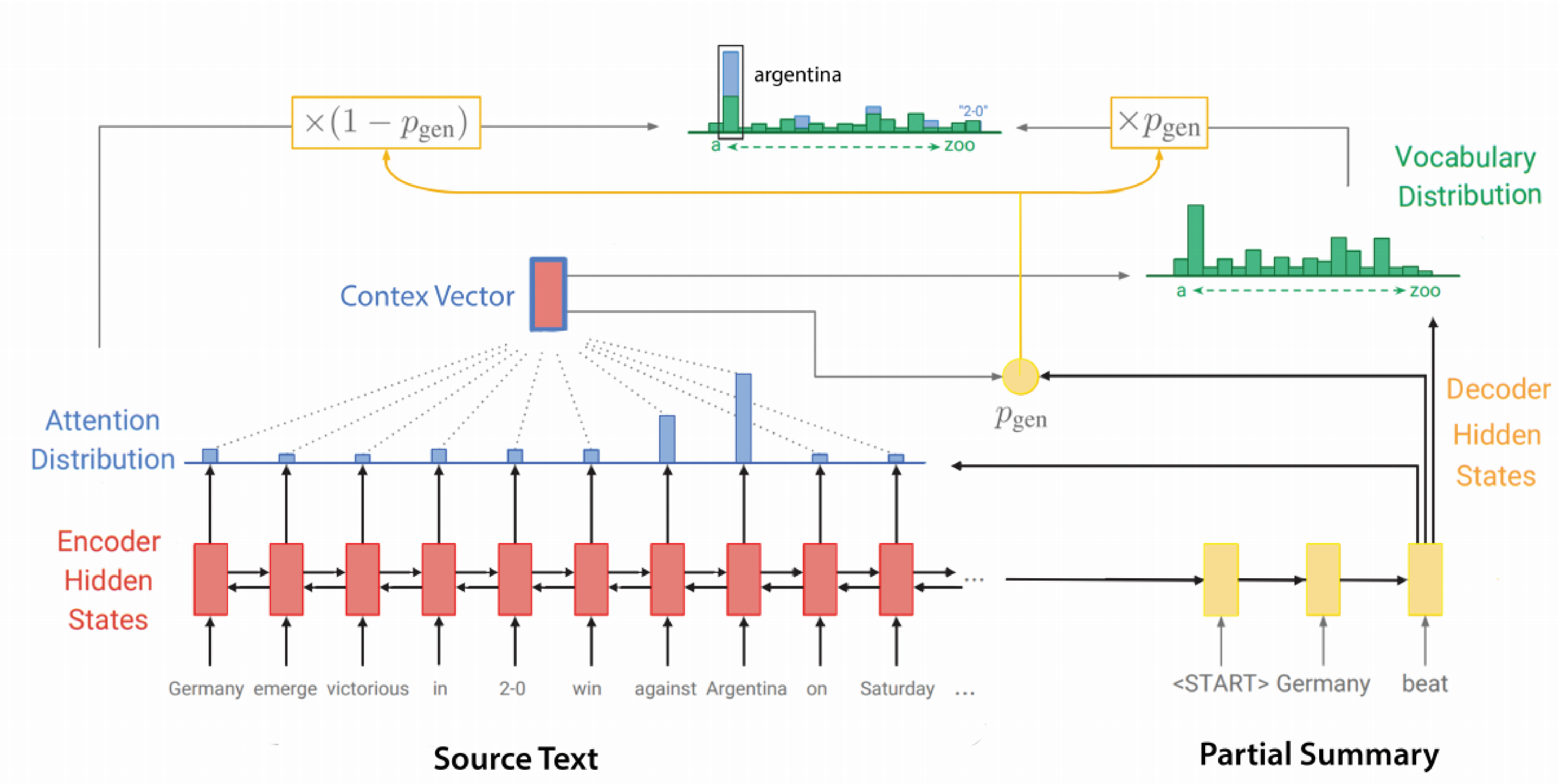
Sequence-to-sequence with attention



Sequence-to-sequence with attention



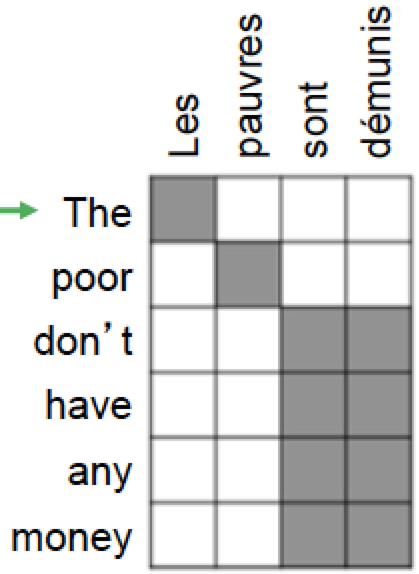
RNN with attention mechanisms



Source: See, Liu, Manning - 2017

Attention is great

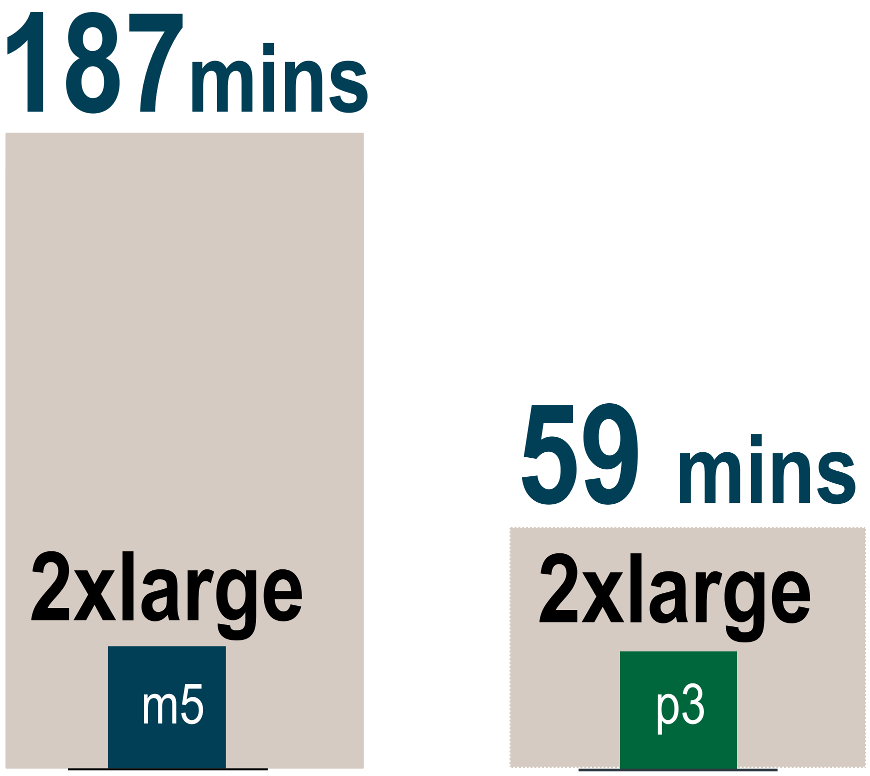
- Attention significantly improves NMT performance
  - It’s very useful to allow decoder to focus on certain parts of the source
- Attention solves the bottleneck problem
  - Attention allows decoder to look directly at source; bypass bottleneck
- Attention helps with vanishing gradient problem
  - Provides shortcut to faraway states
- Attention provides some interpretability
  - By inspecting attention distribution, we can see what the decoder was focusing on
  - We get alignment for free!
  - This is cool because we never explicitly trained an alignment system
  - The network just learned alignment by itself



"Abstracts" from the model:

TEXT:
"great taffy at a great price. there was a wide assortment of yummy taffy. delivery was very quick. if your a taffy lover, this is a deal."
PREDICTED SUMMARY:
nice taffy!
ACTUAL SUMMARY:
great taffy!

The power of P3 Instance on 50K items



Let's go build!

AI SERVICES

Vision			Speech		Language		Chatbots	Forecasting	Recommendations
REKOGNITION IMAGE	REKOGNITION VIDEO	TEXTRACT	POLLY	TRANSCRIBE	TRANSLATE	COMPREHEND	LEX	FORECAST	PERSONALIZE

ML SERVICES

AMAZON SAGEMAKER

BUILD	TRAIN	DEPLOY
Pre-built algorithms & notebooks	One-click model training & tuning	One-click deployment & hosting
Data labeling (GROUND TRUTH)	Optimization (NEO)	
Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)	Reinforcement learning	

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure

EC2 P3 & P3N

EC2 C5

FPGAs

GREENGRASS

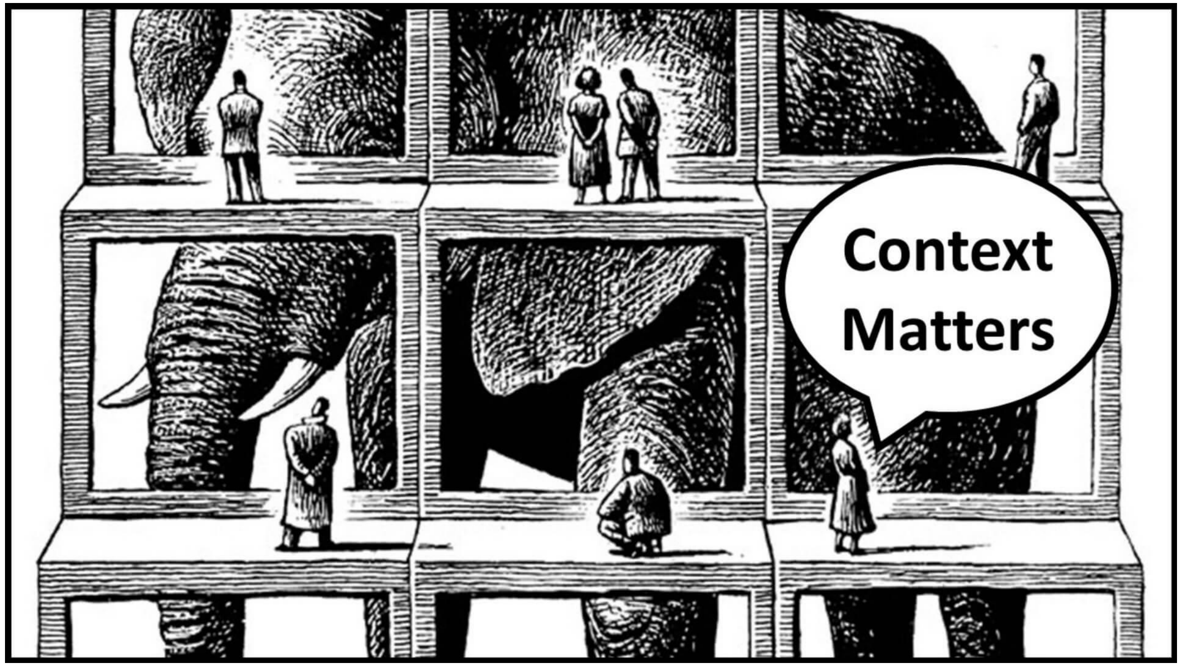
ELASTIC INFERENCE

aws

Thank you for your interest.

The Goal: Pre-train + Finetune in NLP

Previously, context representation was either one directional, or only token level (missing the bigger picture)



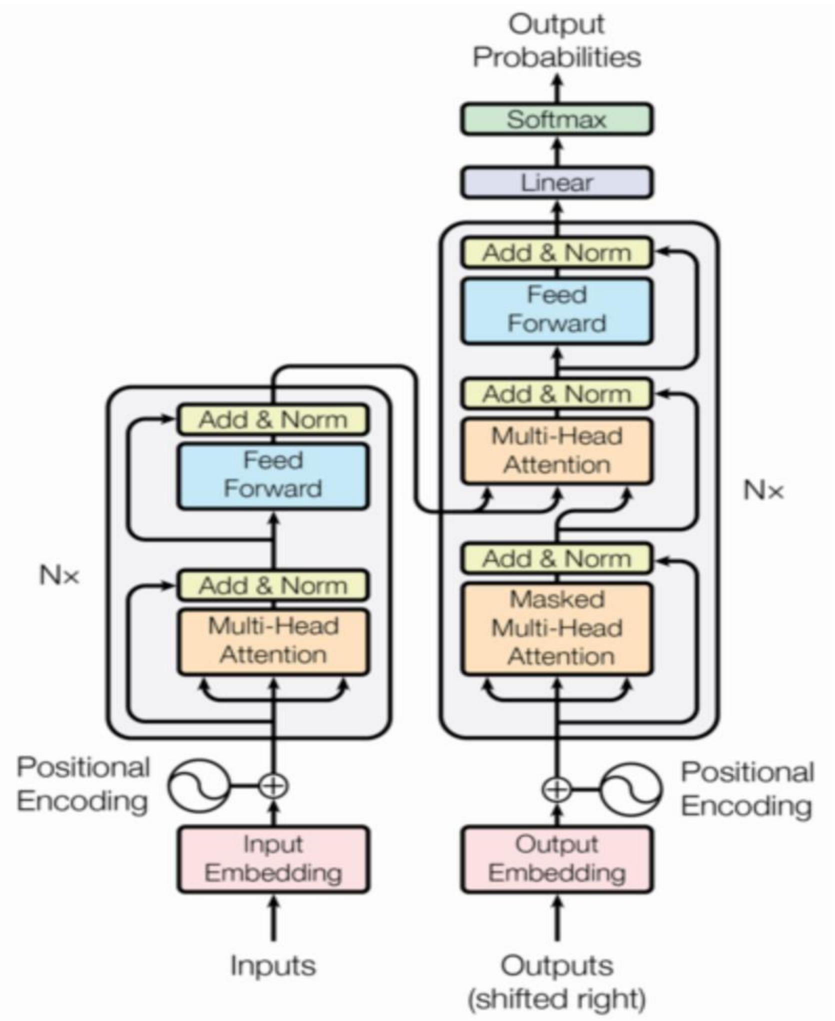
2018 Major NLP Advances

- Transformer – Attention Is All You Need
  - Vaswani et al. (Google) technically 2017
- ULMFiT – Universal Language Model Fine-tuning for Text Classification
  - Howard & Ruder (fast.ai, AYLIEN)
- ELMo – Deep contextualized word representations
  - Peters et al. (AI2, UW)
- GPT Transformer – Improving Language Understanding by Generative Pre-Training
  - Radford et al. (OpenAI)
- BERT – Pre-training of Deep Bidirectional Transformers for Language Understanding
  - Devlin et al. (Google)

Among many more…

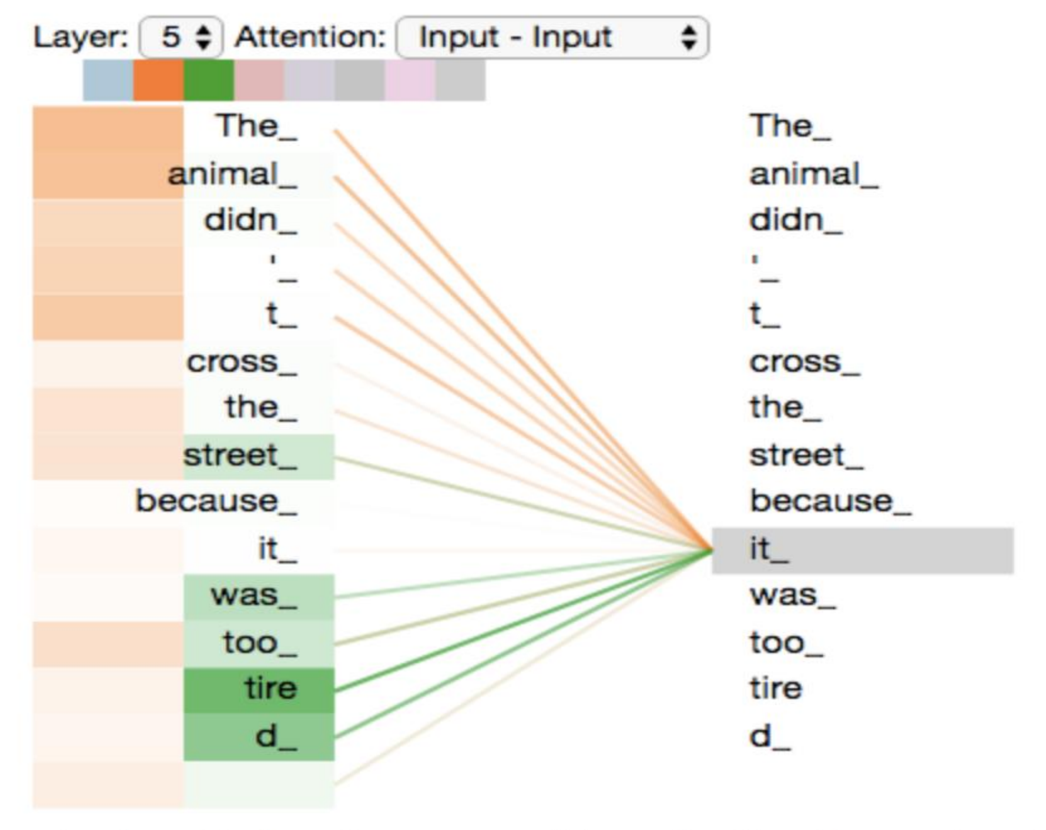
Transformer – Attention Is All You Need

- No recurrent layers (RNN/LSTM); allows parallelization
- Transformer: Basic building block comprises of Attention and FFN layers
- Both Encoder and Decoders comprised of stacked Transformers.
- Can be trained significantly faster.

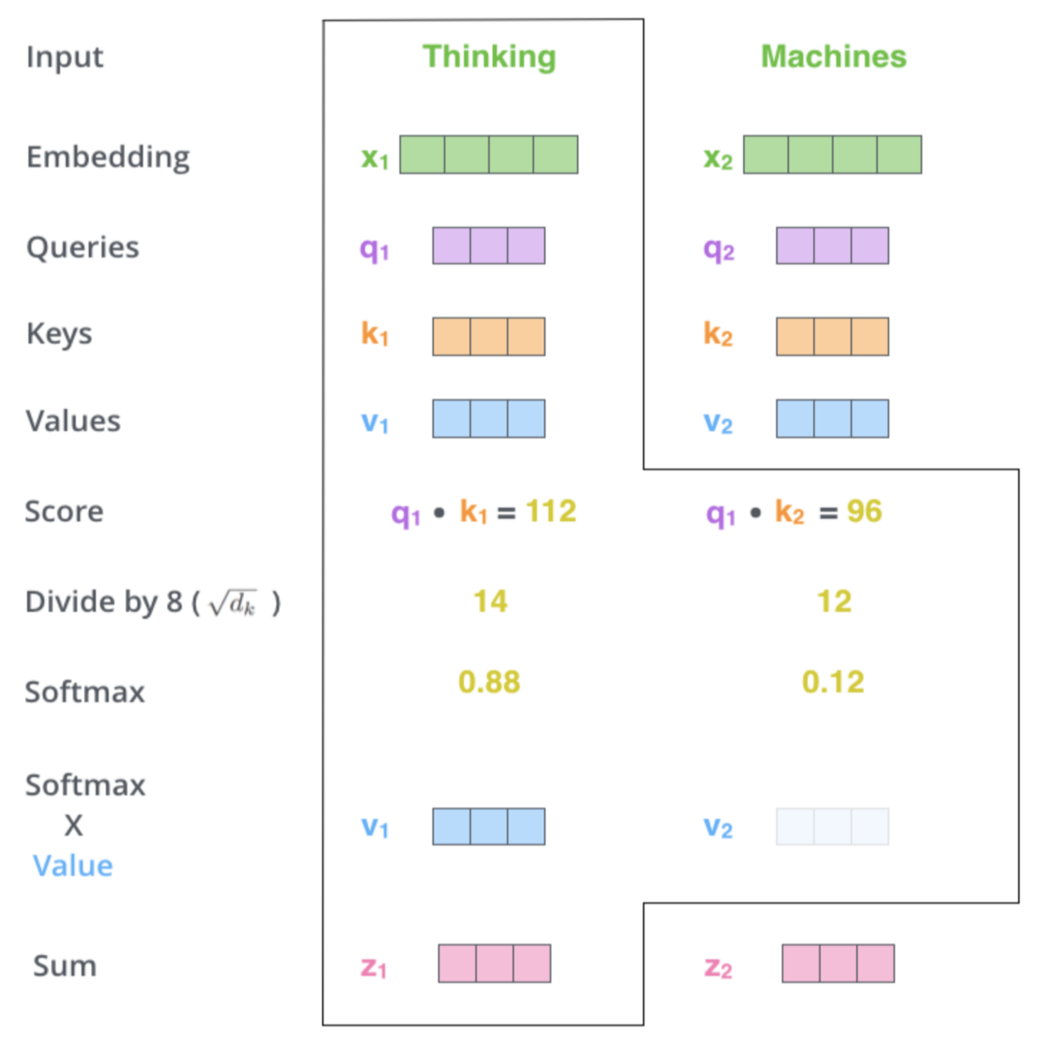


Self-Attention

Intuition



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Credit: https://jalammar.github.io/illustrated-transformer/

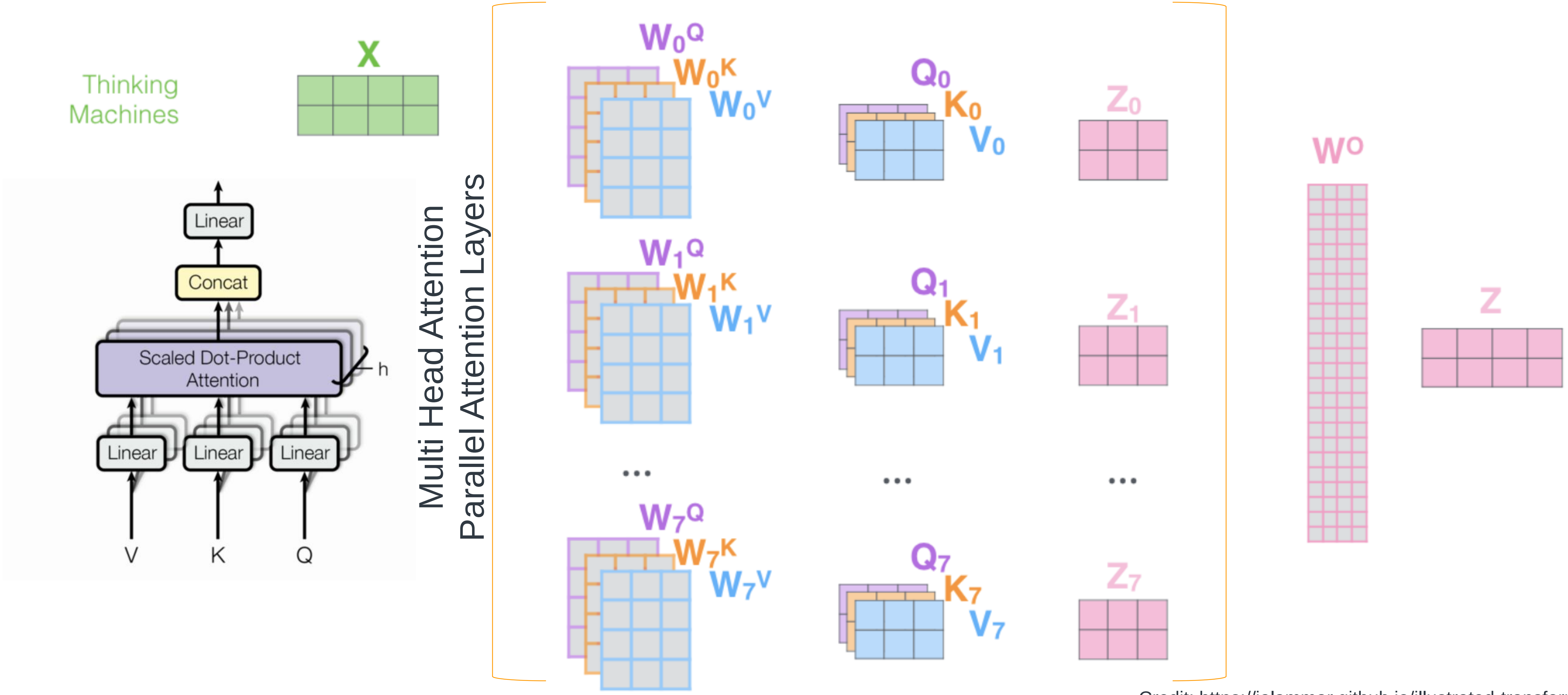
1) This is our input sentence*

2) We embed each word*

3) Split into 8 heads. We multiply X or R with weight matrices

4) Calculate attention using the resulting Q/K/V matrices

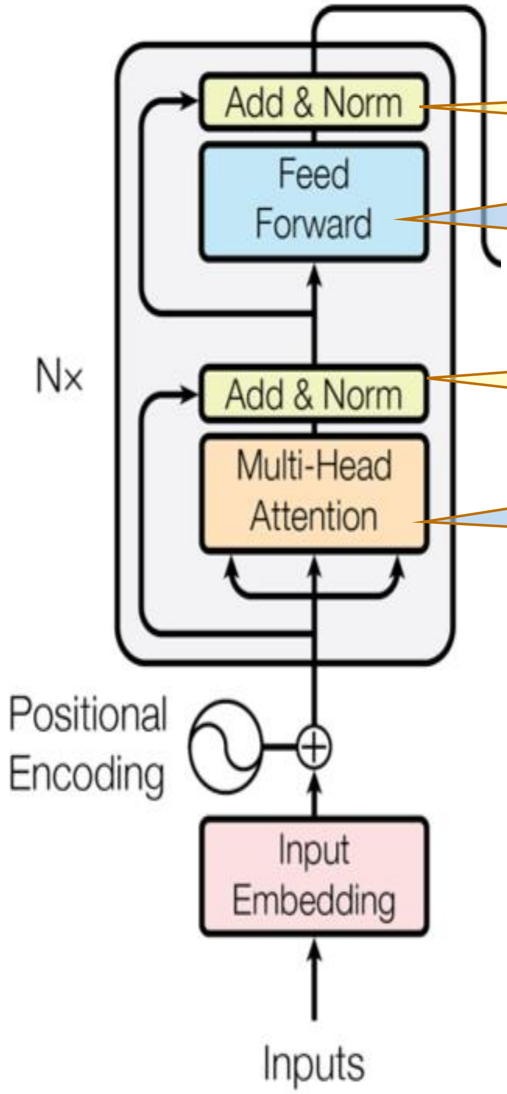
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer



Credit: https://jalammar.github.io/illustrated-transformer/

Encoder

- Constant layer dimension: $d_{model} = 512$
- Employs dropout to every sub-layer before norm and embedding layers

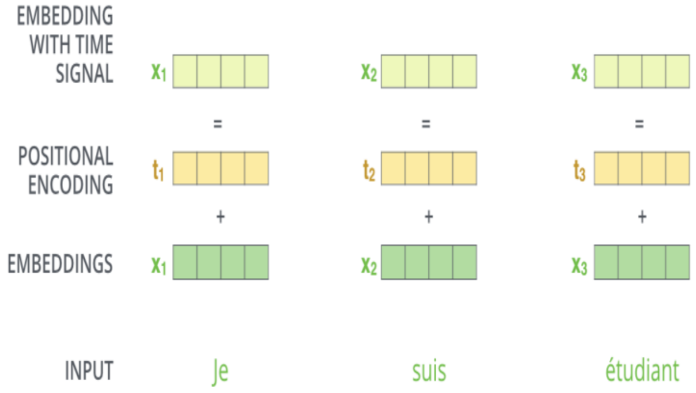


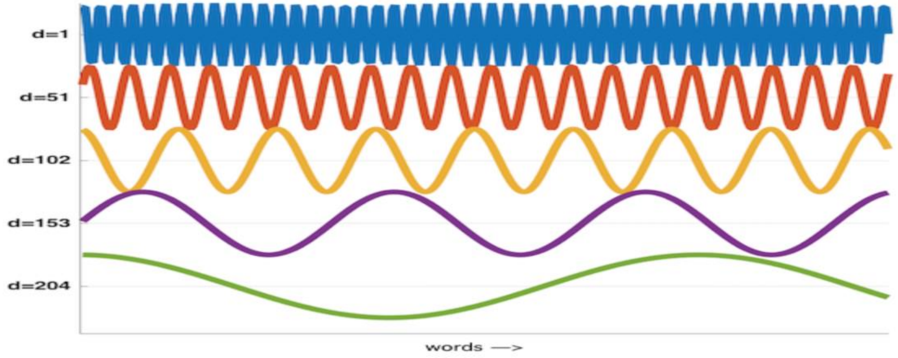
$$Y = LayerNorm(u + FFN(u))$$

$$FFN(x) = Relu(x, W_1)W_2 + b$$
$$W_1 \in \mathcal{R}^{d_{model} \times 2048}, \quad W_2 \in \mathcal{R}^{2048 \times d_{model}}$$

$$Y = LayerNorm(u + Multi\text{-}Head\text{-}Attn(u))$$

Comprises of 8 Self-Attention layers

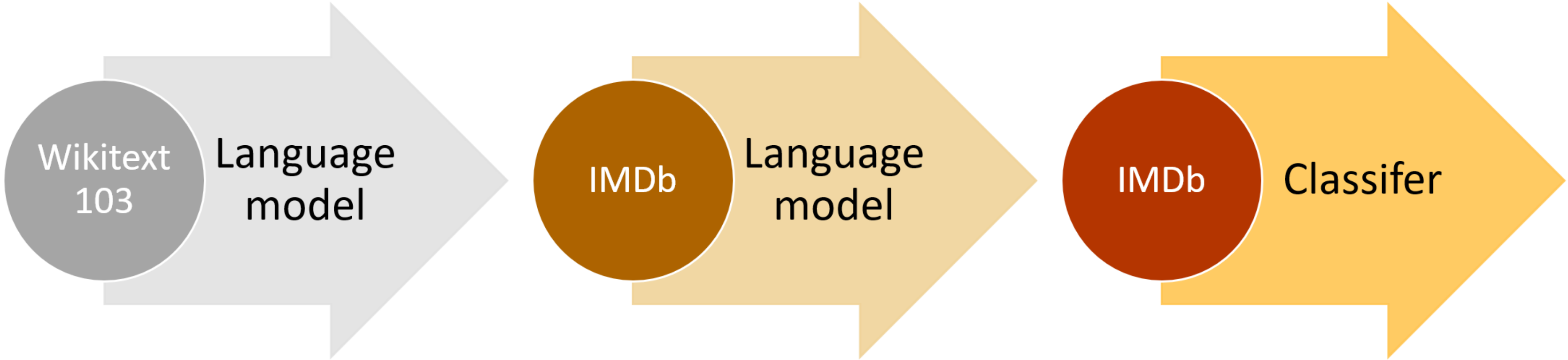




$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

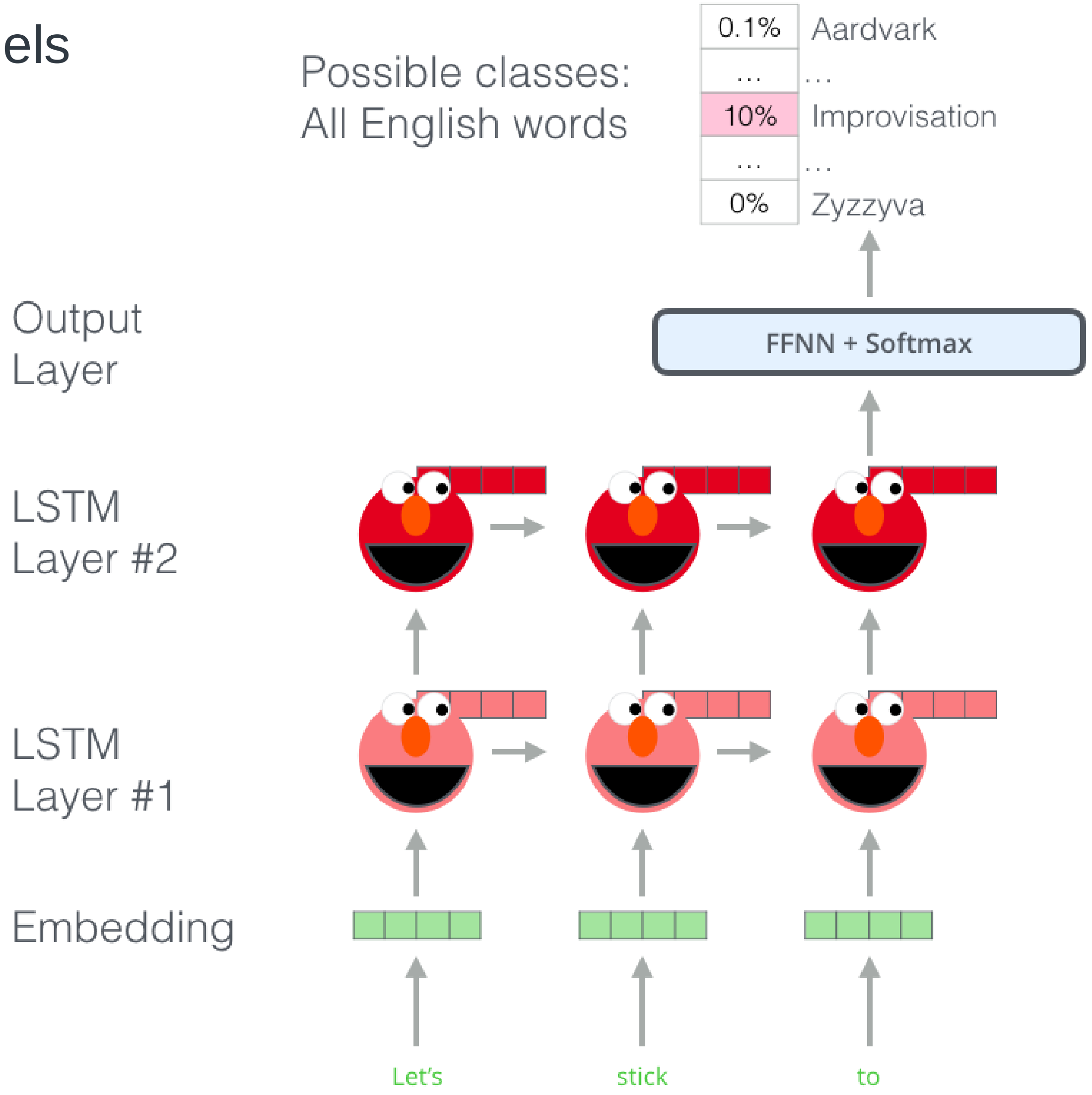
ULMFiT – Universal Language Model Fine-tuning for Text Classification

- Key takeaways:
  - Effective transfer learning for NLP (using LSTMs)
  - Introduces novel language model fine-tuning techniques
  - Helps solve NLP problems with less data



ELMo – Embeddings from language models

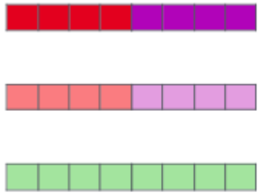
- Key takeaways:
  - Word embedding values conditioned on context
    - Handles polysemy
  - Trained using BiLSTM on next-word-prediction task



ELMo – Deep contextualized word representations

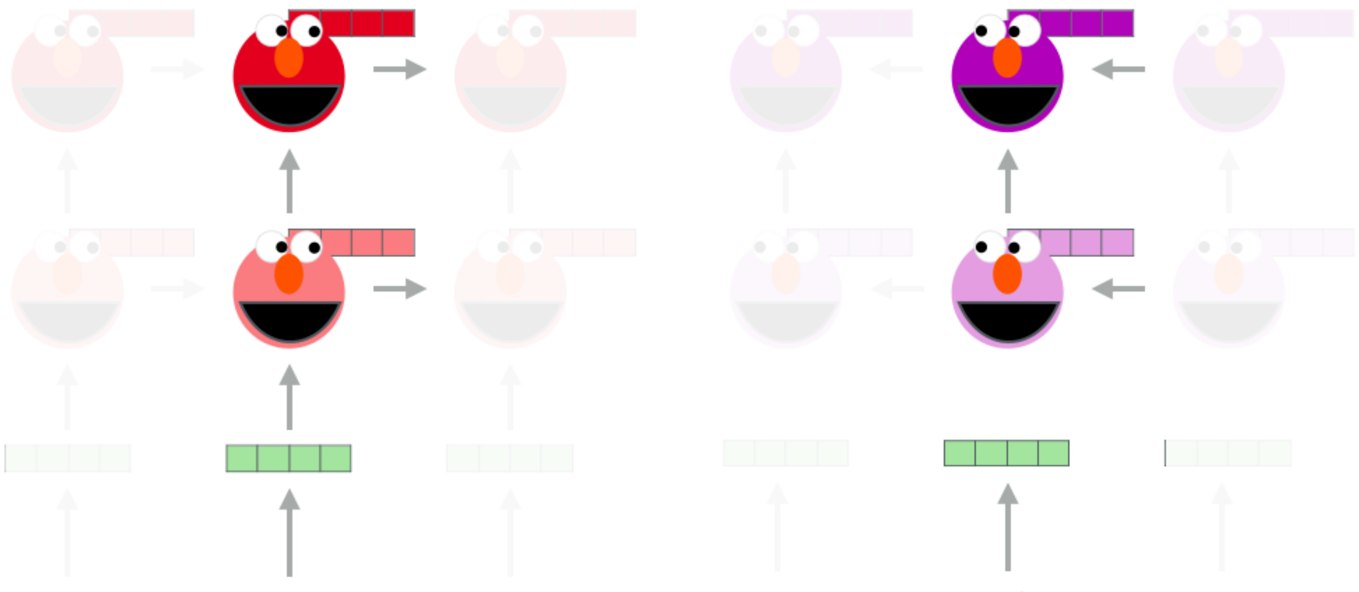
Embedding of “stick” in “Let’s stick to” - Step #2

1- Concatenate hidden layers



Forward Language Model

Backward Language Model



2- Multiply each vector by a weight based on the task

x s_2

x s_1

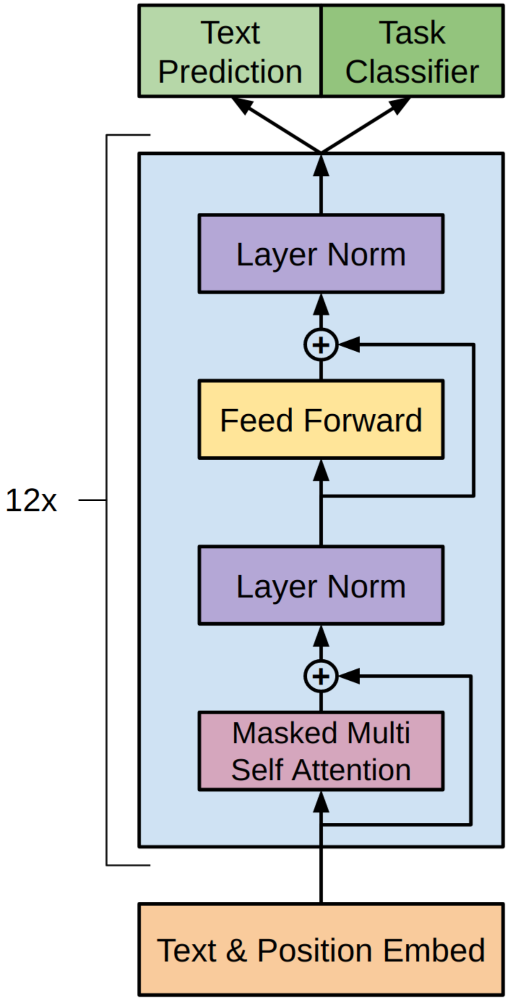
x s_0

3- Sum the (now weighted) vectors

ELMo embedding of “stick” for this task in this context

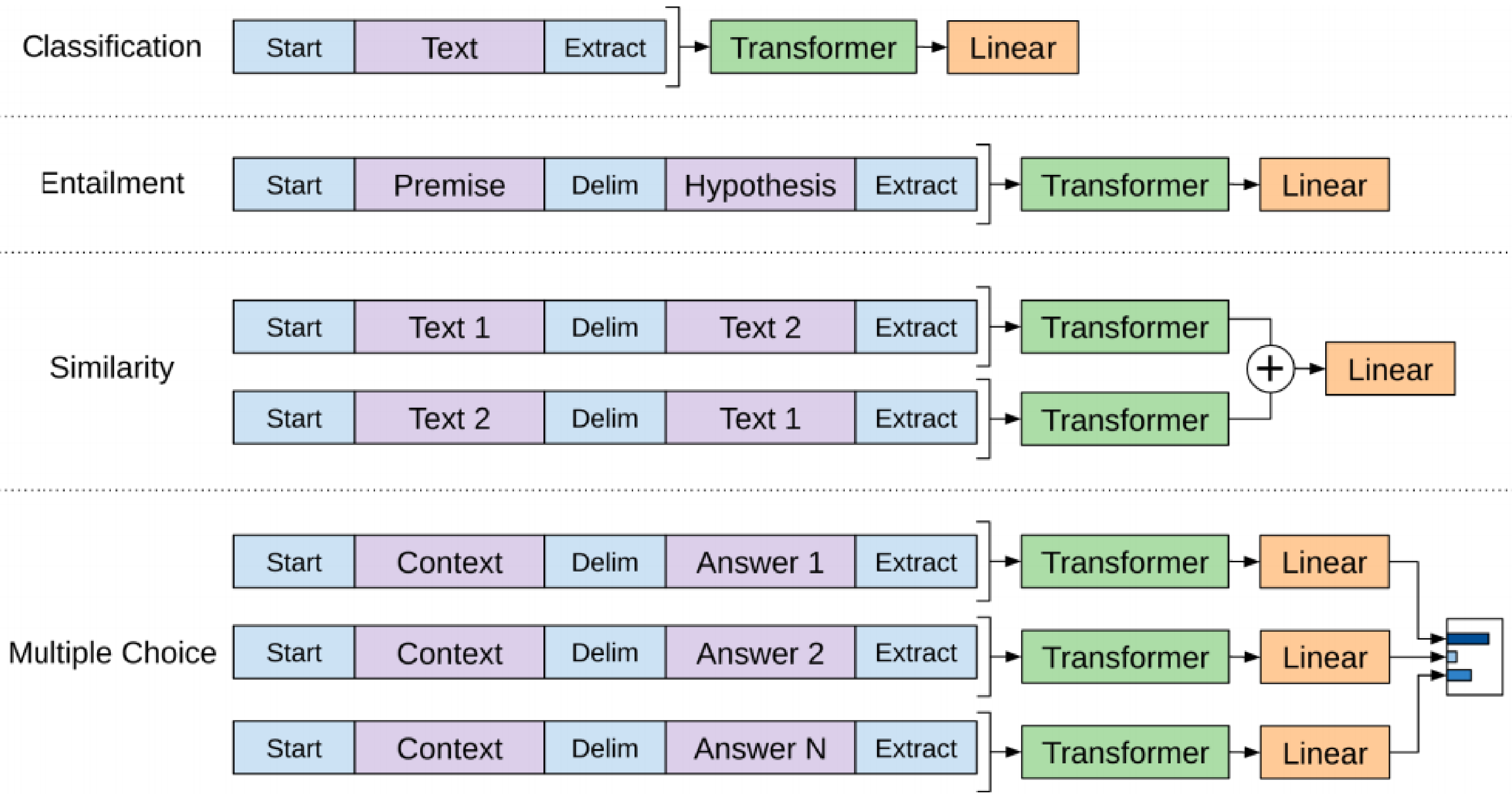
GPT Transformer – Generative Pre-Training

- Setting the stage for multi-task NLP
- Key takeaways:
  - Combining unsupervised pre-training with Transformers
    - Building upon ULMFiT, ELMo
  - The OpenAI Transformer
    - **Only Transformer decoders**, trained on prediction and classification
    - No encoder-decoder attention sublayer
    - Remember: Transformer decoder masks future tokens
      - Note: Only a forward language model, **not bidirectional**
  - SOTA performance on GLUE benchmark
    - Shows Transformer is flexible and robust



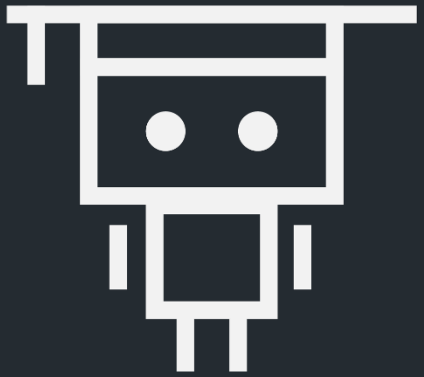
GPT Transformer – Generative Pre-Training

- Multitasking trick: Input transformations for various tasks

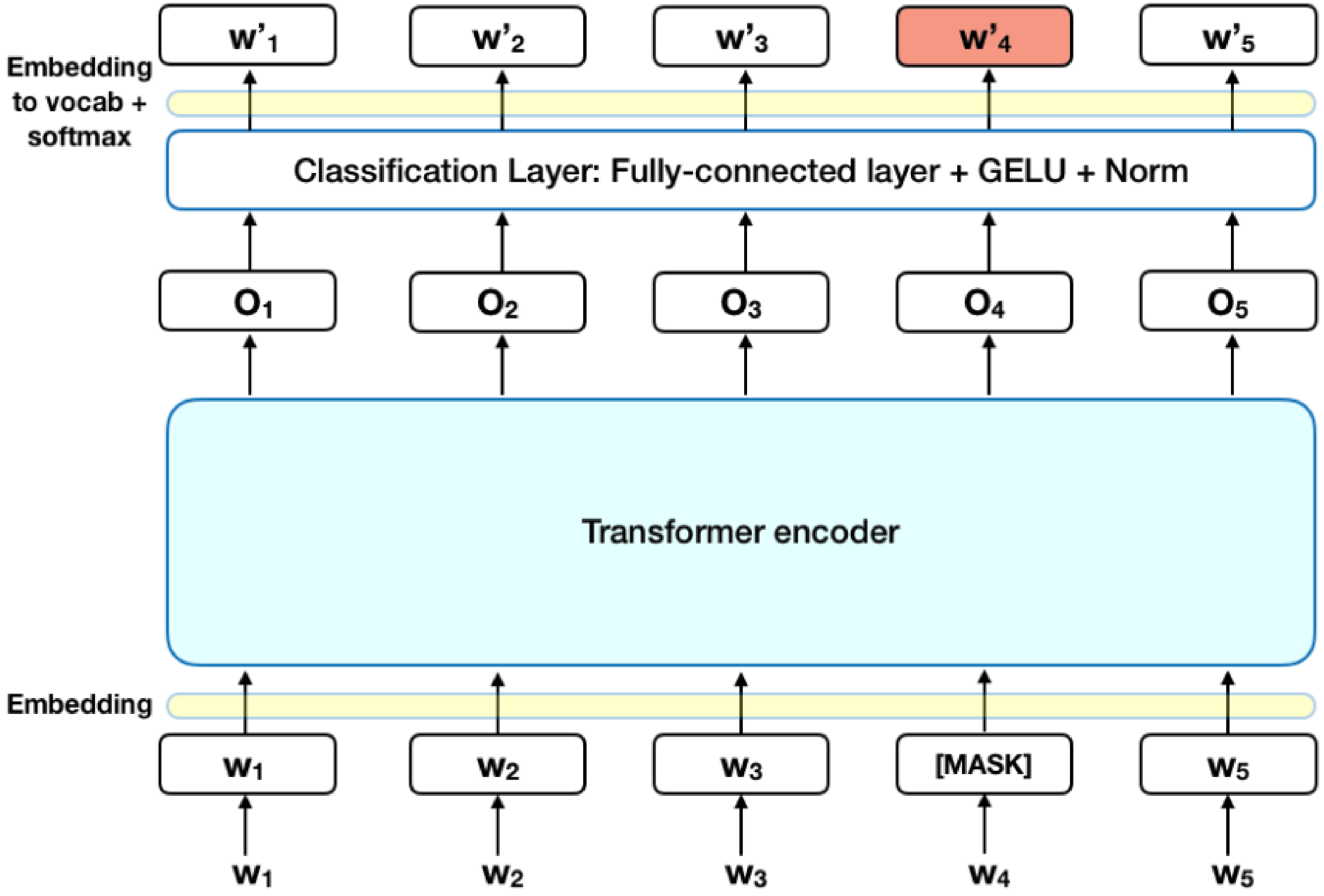


BERT:

Bidirectional Encoder Representations from Transformers

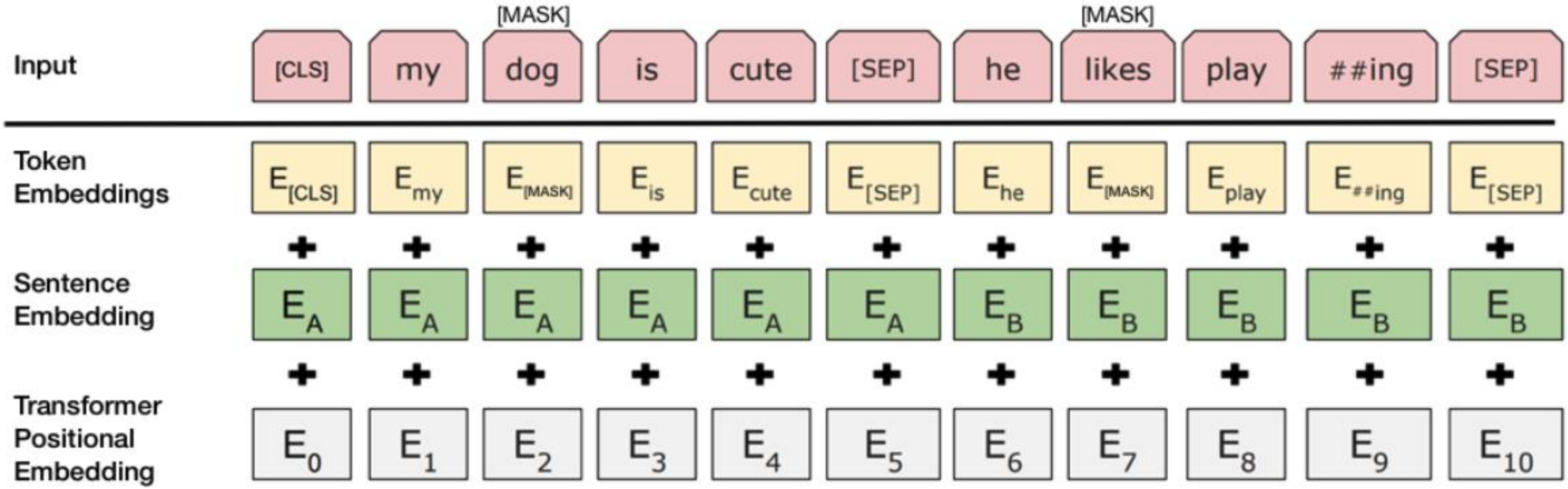


Secret Sauce #1: Masked LM

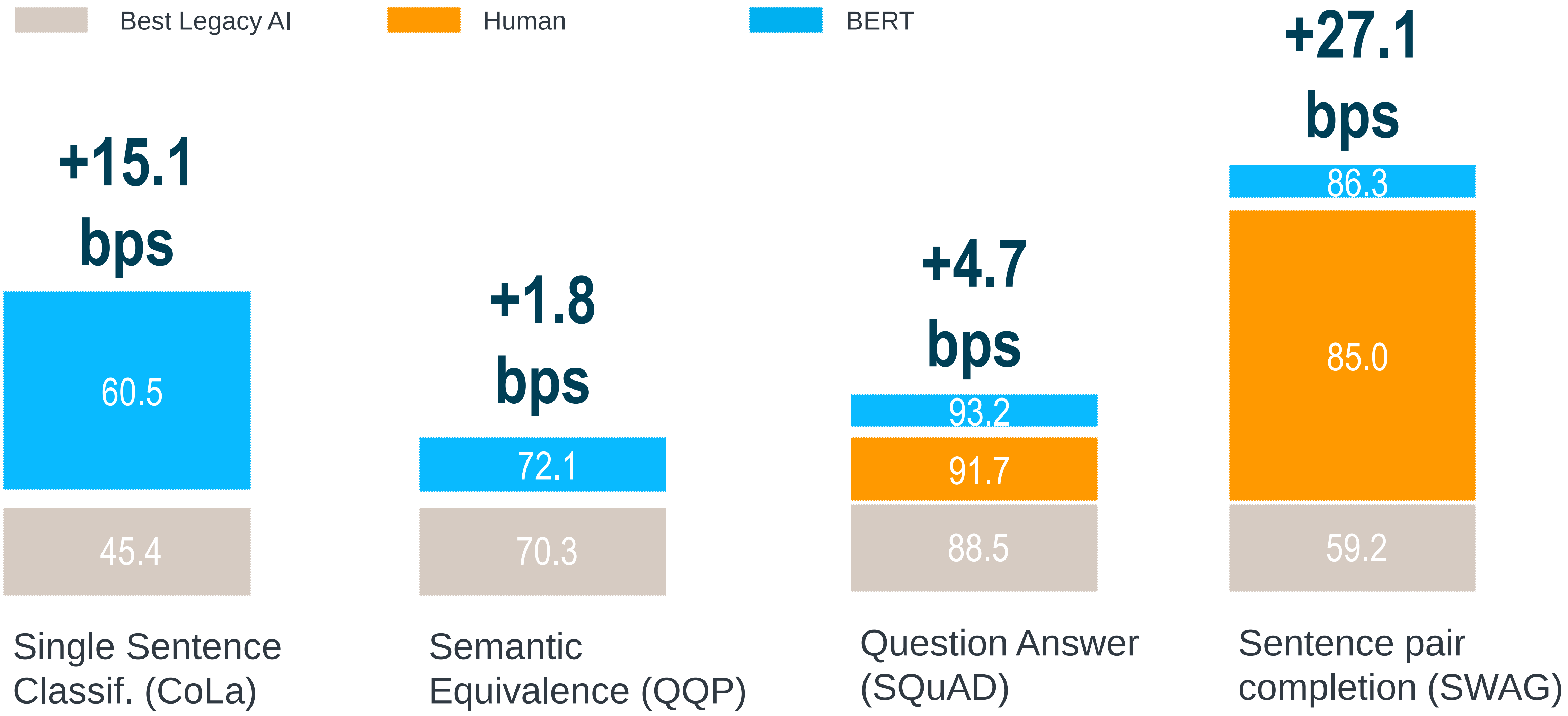


- Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token
- The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence

Secret Sauce #2: Next Sent. Pred.



Results: Surpassing Humans



Thank you!

amazon

MACHINE LEARNING UNIVERSITY