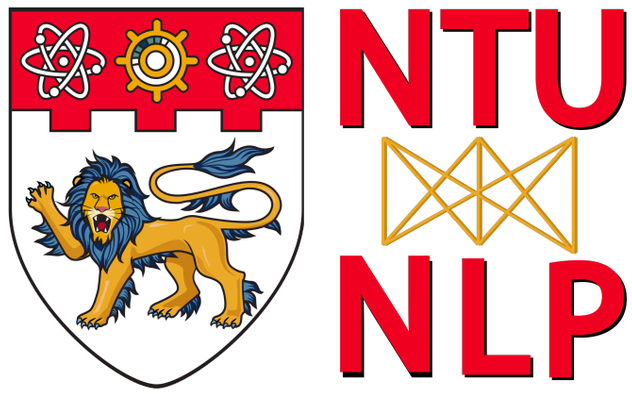
Deep Learning for Natural Language Processing

Shafiq Joty



Lecture 8: Seq2Seq and Transformers

About me

Home

Papers

Research

Teaching

Students

CV

Shafiq Rayhan Joty

Associate Professor (tenured)
NTU Natural Language Processing Group
Office: Block N4, 02c-79
School of Computer Science and Engineering
Nanyang Technological University, Singapore

Email
Phone
Google Scholar
Github

Senior Research Manager
Salesforce AI Research
Office: Floor 40, Suntec City Tower 2

Email
Phone

Short Biography

- Associate Professor (with tenure), Nanyang Technological University (NTU), Singapore [Mar'22 -]
- Senior Research Manager, Salesforce AI Research, Singapore [Feb'19 -]
- Assistant Professor, Nanyang Technological University (NTU), Singapore [Jul'17 - Feb'22]
- Research Scientist, Qatar Computing Research Institute (QCRI), Doha [Jan'14 - Jul'17]
- PhD in Computer Science, University of British Columbia (UBC), Vancouver [Sep'08 - Feb'14]

Research Interests

Natural Language Processing

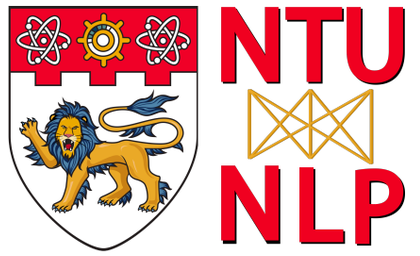
- NLP tools (Syntax, Semantics and Discourse)
- Multi-lingual NLP (Machine Translation, Cross-lingual tasks)
- NLP Applications (QA, Summarization, Dialogue)
- Multi-modal NLP (Image/Video Captioning)
- Robust/adversarial NLP
- NLP for Programming

Machine Learning

- Deep Learning
- Probabilistic Graphical Models
- Reinforcement Learning

https://raihanjoty.github.io/

Where we are



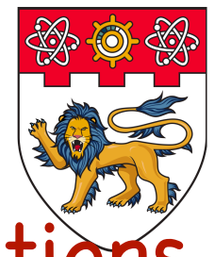
Models/Algorithms

- Linear models
- Feed-forward Neural Nets (FNN)
- Window-based methods
- Convolutional Nets
- Recurrent Neural Nets

NLP tasks/applications

- Word meaning
- Language modelling
- Sequence tagging
- Sequence encoding

Plan for 2nd Half



NTU NLP

Models/Algorithms

- Seq2Seq variants (Wk 8)
- Transformers (Wk 8 + 9)
- Self-supervised learning (Wk 9 + 10)
- Multilingual NLP (Wk 10 + 11)
- Adversarial NLP (Wk 12)

NLP tasks/applications

- Language modelling
- Sequence tagging
- Sequence encoding
- Machine Translation
- Summarization
- …….

Today

NTU NLP

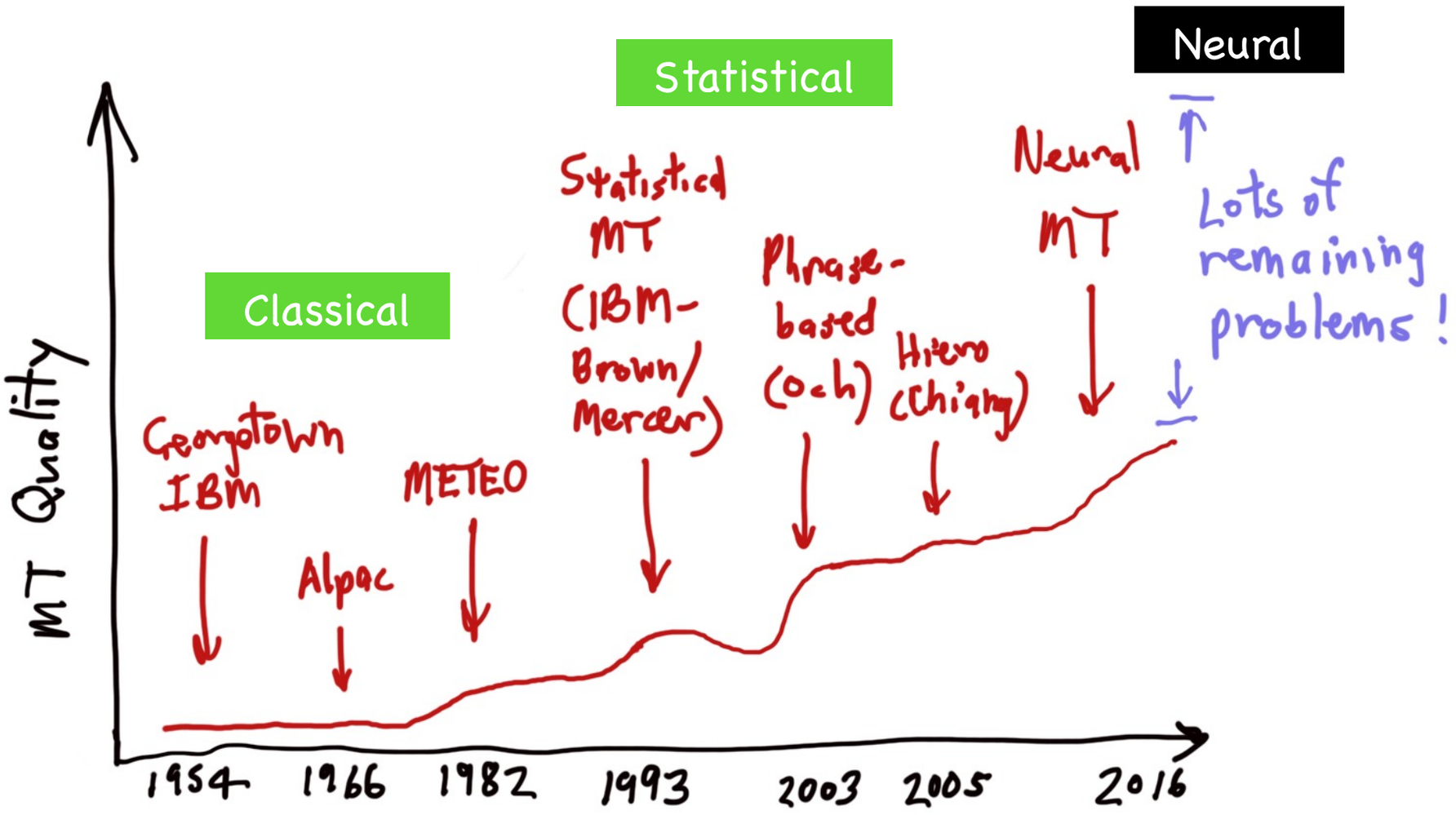
Models/Algorithms

- Seq2Seq
  - + Attention (**Recap**)
  - + Subword
- **Seq2Seq Varints**
- **Transformer Seq2Seq**

NLP tasks/applications

- **Machine Translation**
- **Summarization**
- **Parsing**
- **Dialogue generation**

Progress in MT



- Microsoft, Google, Yandex claimed **human parity** in MT in 2018 with NMT. Only true in a controlled setup

Source: Luong, Cho, Manning (ACL tutorial 2016)

Neural Machine Translation (2014 -)

NTU NLP

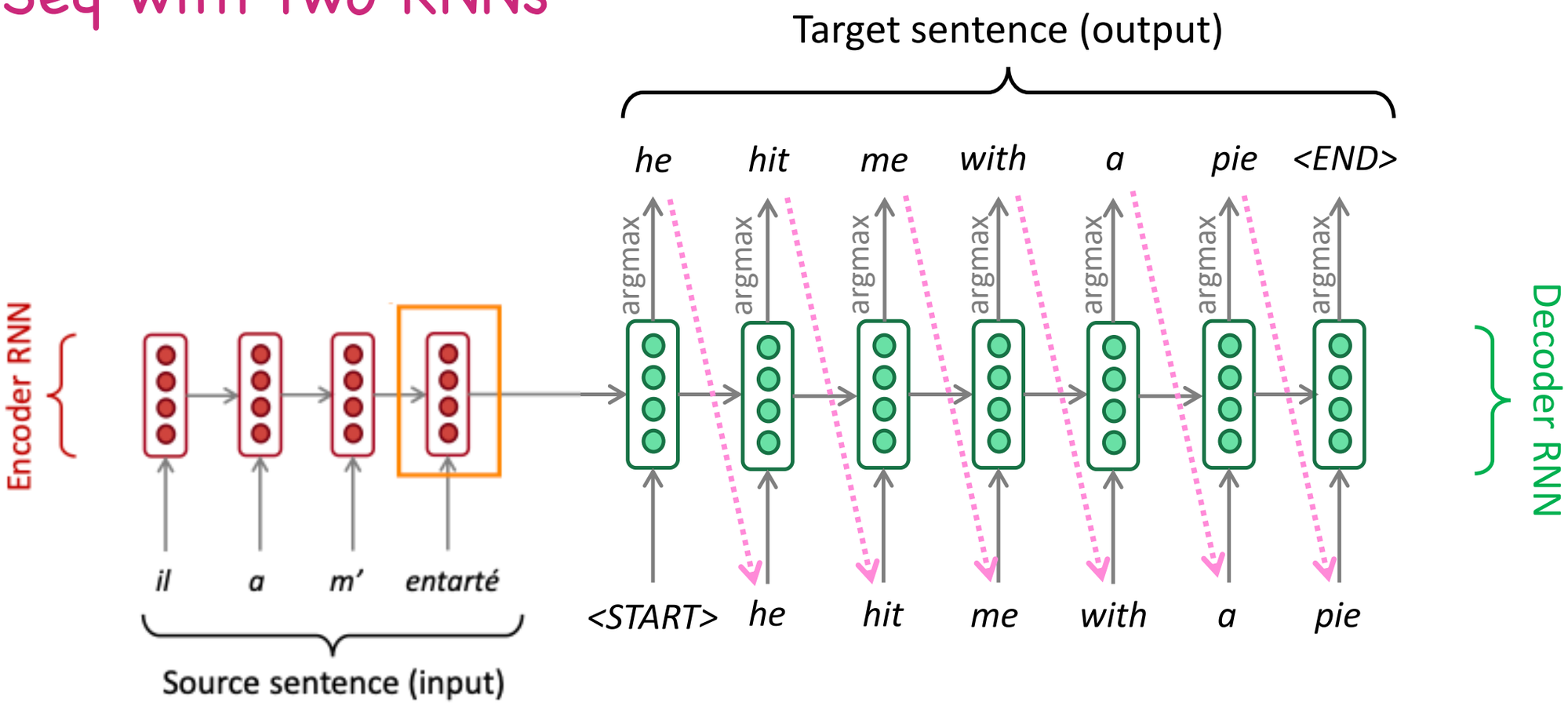
- New paradigm

- One Network for everything!

Photo credit: Internet

Sequence-to-sequence NMT Model

- Seq2Seq with Two RNNs

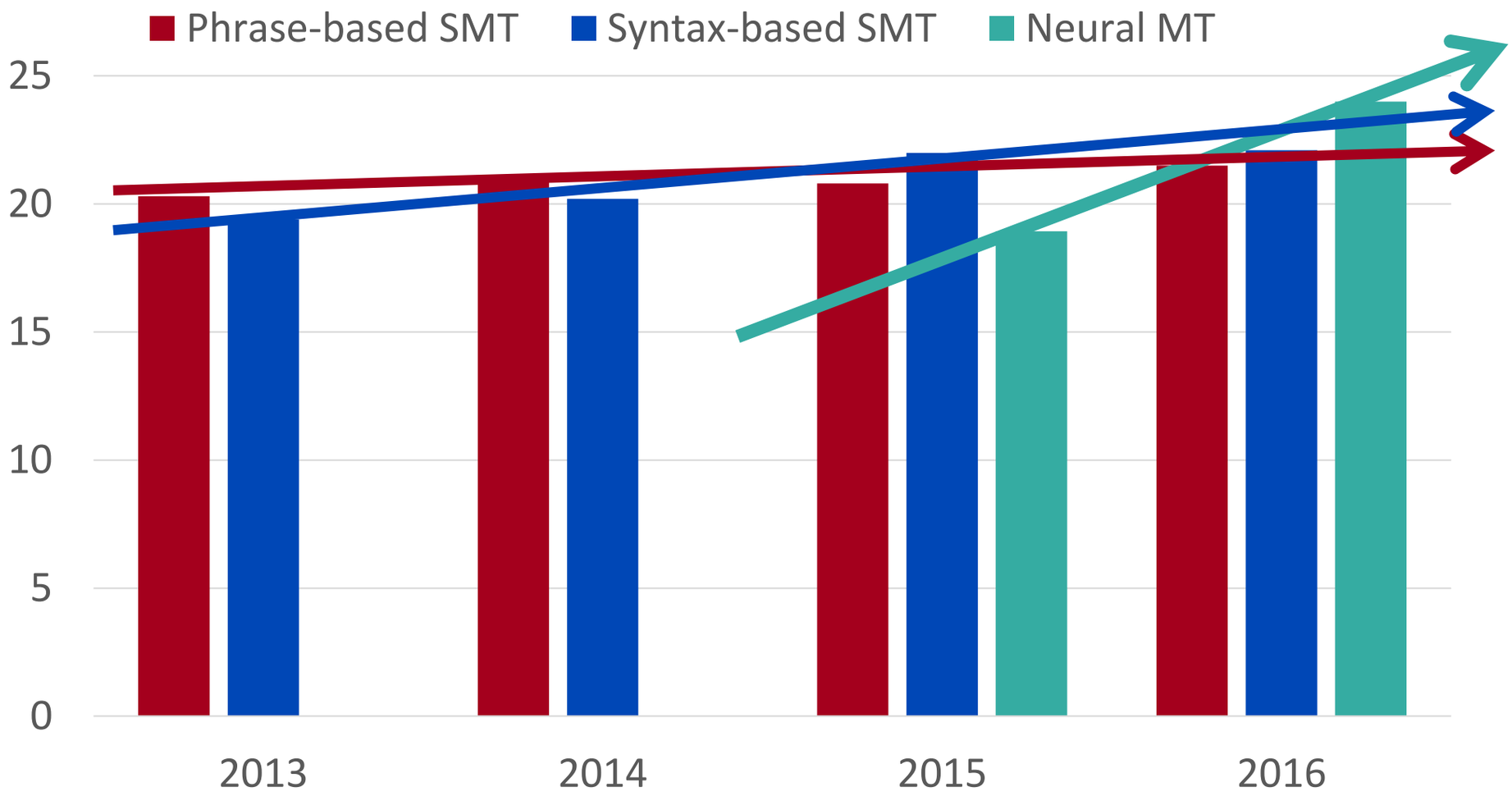


Encoder produces an encoding of the source sentence.

Provides Initial hidden state for Decoder

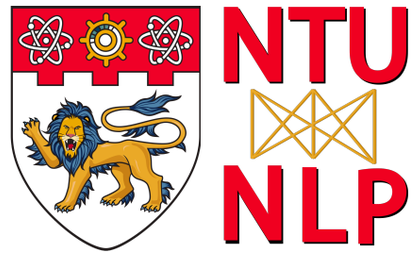
Decoder RNN is a **Conditional Language Model** that generates target sentence, conditioned on encoding.

NMT Success



Source: http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf

NMT Success



NMT: the biggest success story of NLP Deep Learning

- Neural Machine Translation went from a fringe research activity in **2014** to the leading standard method in **2016**
- **2014**: First seq2seq paper published [1]
- **2016**: Google Translate switches from SMT to NMT
- This is amazing! **SMT** systems, built by thousands of engineers over many years, outperformed by NMT systems trained by a handful of engineers in a few months

[1] Sequence to Sequence Learning with Neural Networks. Ilya Sutskever, Oriol Vinyals, Quoc V. Le. NIPS 2014

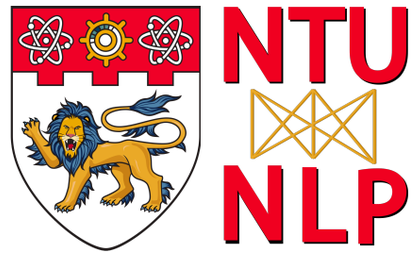
Is MT Solved?

- Is MT solved?
  - Nope!
  - Although there are claims from Google, Microsoft, Yandex that their MT systems have achieved human parity, this is true for only constrained setups [1,2].

[1] Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In EMNLP, 2018.

[2] M. Popel, M. Tomková, J. Tomek, Lukasz Kaiser, Jakob Uszkoreit, Ondrej Bojar, and Z. Žabokrt-ský. Transforming machine translation: a deep learning system reaches news translation qualitycomparable to human professionals. Nature Communications, 11, 2020.

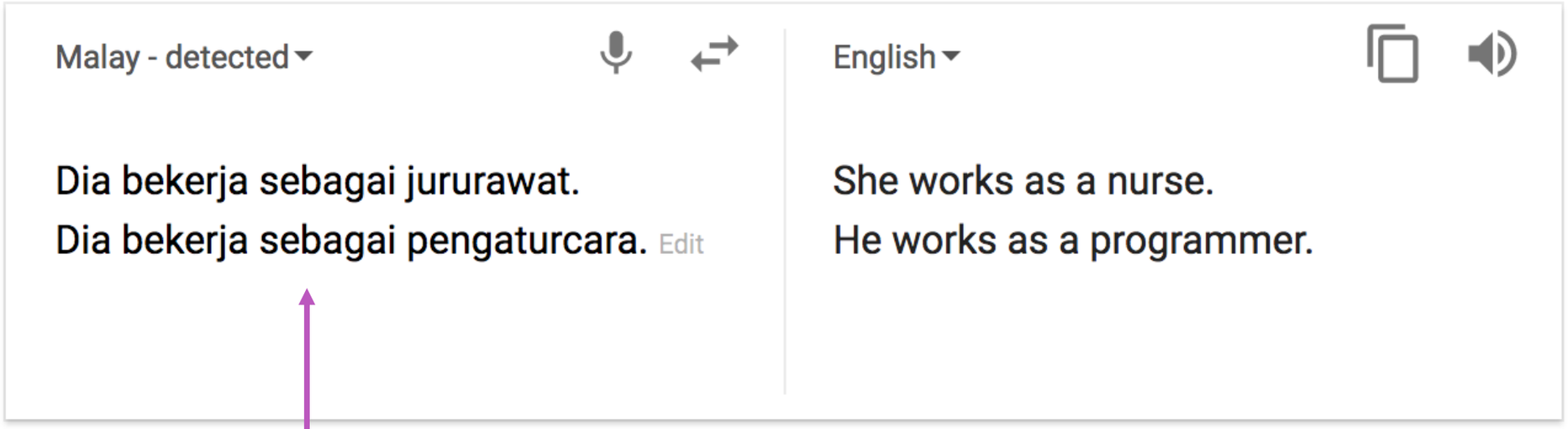
Is MT Solved?



- Still many difficulties:
  - Data hungry! Has been successful for high-resource languages
    - How to deal with low-resource languages (or domains)?
    - Most languages are low-resource
  - Dealing with out-of-vocabulary words
    - Input and Output dictionaries need to be fixed and limited
  - Maintaining longer context
    - Discourse-level aspects (anaphora, connectives)
  - Domain mismatch between train and test data
  - How to interpret
  - Translation speed

Is MT Solved?

- Nope!
- **Dataset biases!**



Source: Stanford 224n

Is MT Solved?

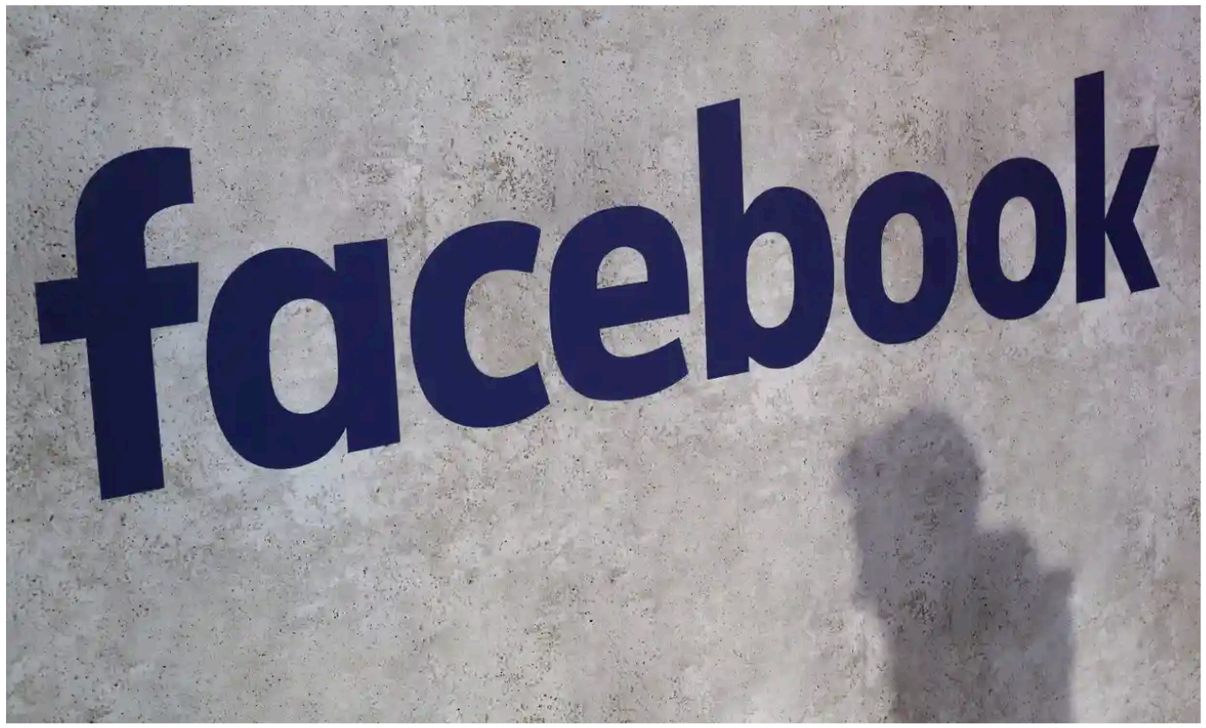
- Nope!
- **Dataset biases (low-resource)!**

Sign in Search International edition

The Guardian

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer



▲ Facebook's machine translation mix-up sees man questioned over innocuous post confused with attack threat. Photograph: Thibault Camus/AP

Is MT Solved?

- Nope!
- OOV word!

BBC Sign in News Sport Reel Worklife Travel Future More

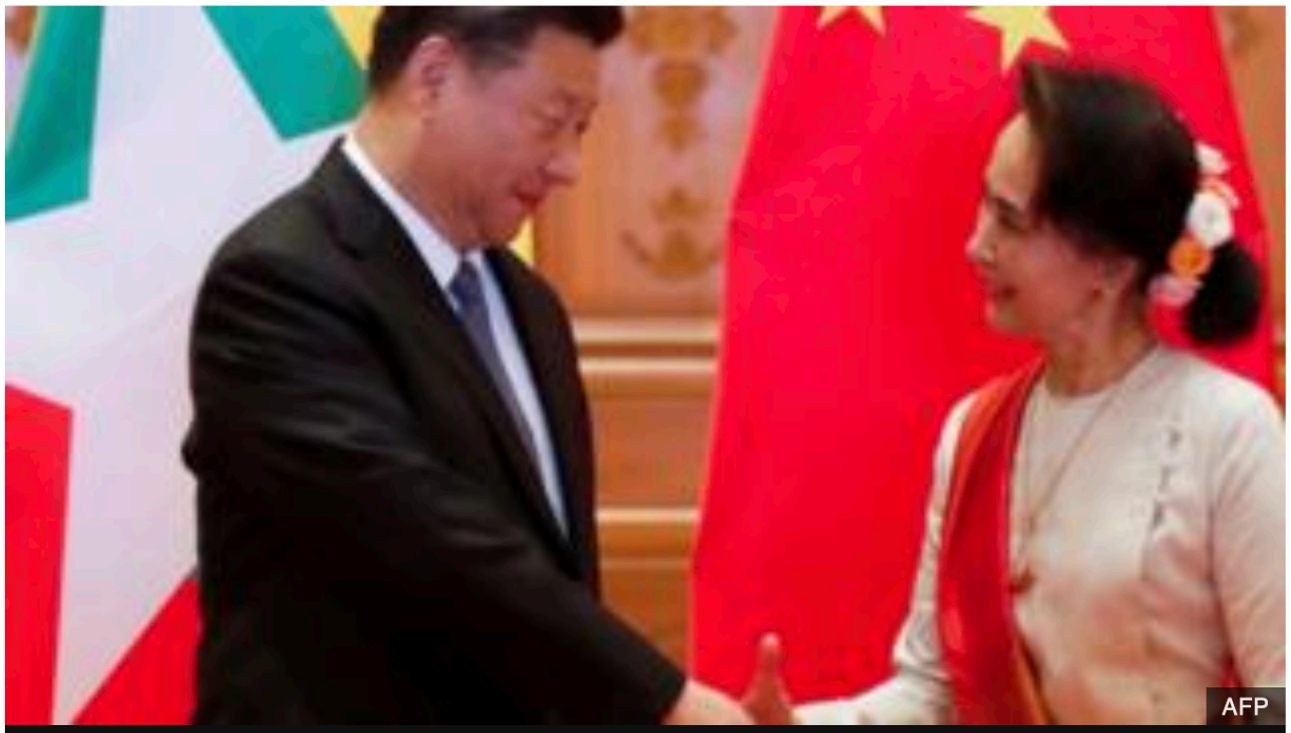
NEWS

Home Video World Asia UK Business Tech Science Stories Entertainment & Arts

Asia China India

Facebook blames 'technical issue' for offensive Xi Jinping translation

19 January 2020 Share



Facebook blamed a "technical issue" for the mistranslation of Xi Jinping's name

Facebook has apologised for translating Chinese President Xi Jinping's name from Burmese to English into an obscenity on its platform.

Is MT Solved?

- Nope!
- Incorporating **common sense knowledge is still hard**

English ▾ | Spanish ▾

paper jam Edit

Mermelada de papel

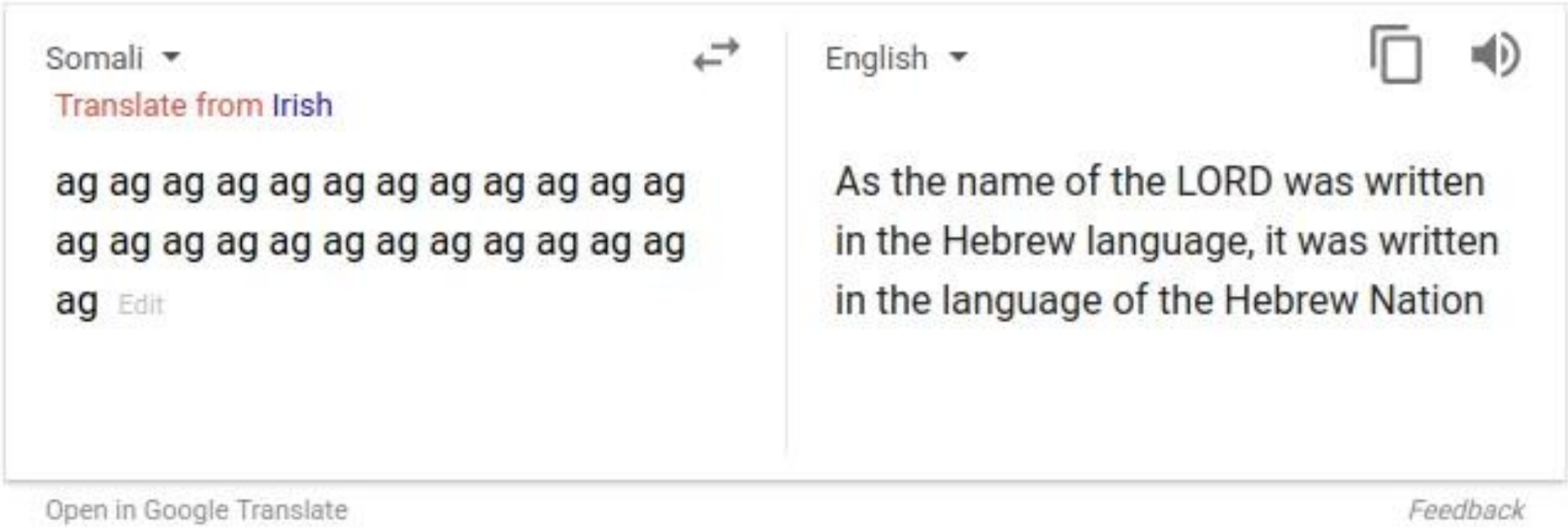
Open in Google Translate | *Feedback*

?

Source: Stanford 224n

Is MT Solved?

- Nope!
- **Uninterpretable systems do strange things**

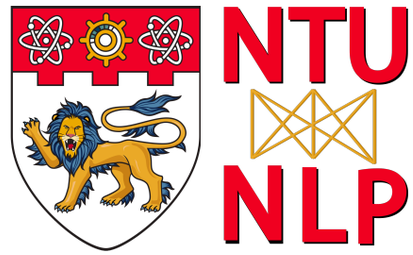


- More examples in Lecture 12

Picture source: https://www.vice.com/en_uk/article/j5npeg/why-is-google-translate-spitting-out-sinister-religious-prophecies

Explanation: https://www.skynettoday.com/briefs/google-nmt-prophecies \

NMT Research Continues!



- NMT has been the **flagship task** for NLP Deep Learning
- NMT research has pioneered many of the recent innovations of Deep Learning (not just in NLP)
- In **2017-2021:** NMT research continues to thrive
- Researchers have found many, many improvements to the "vanilla" seq2seq NMT system we've presented
- Attention is the most important one

Attention Mechanism

Recent Breakthrough

BLOG POST
RESEARCH

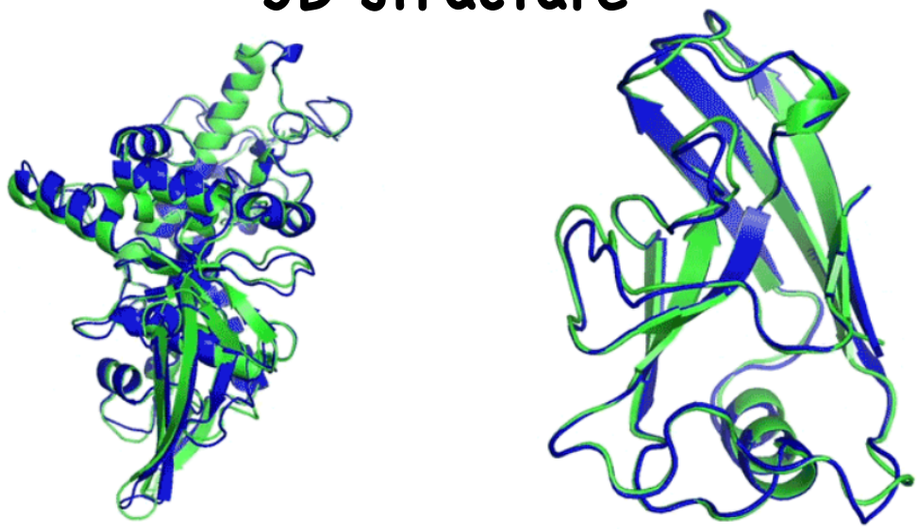
30 NOV 2020

AlphaFold: a solution to a 50-year-old grand challenge in biology

We have been stuck on this one problem – how do proteins fold up – for nearly 50 years. To see DeepMind produce a solution for this, having worked personally on this problem for so long and after so many stops and starts, wondering if we'd ever get there, is a very special moment.

PROFESSOR JOHN MOULT
CO-FOUNDER AND CHAIR OF CASP, UNIVERSITY OF MARYLAND

3D structure



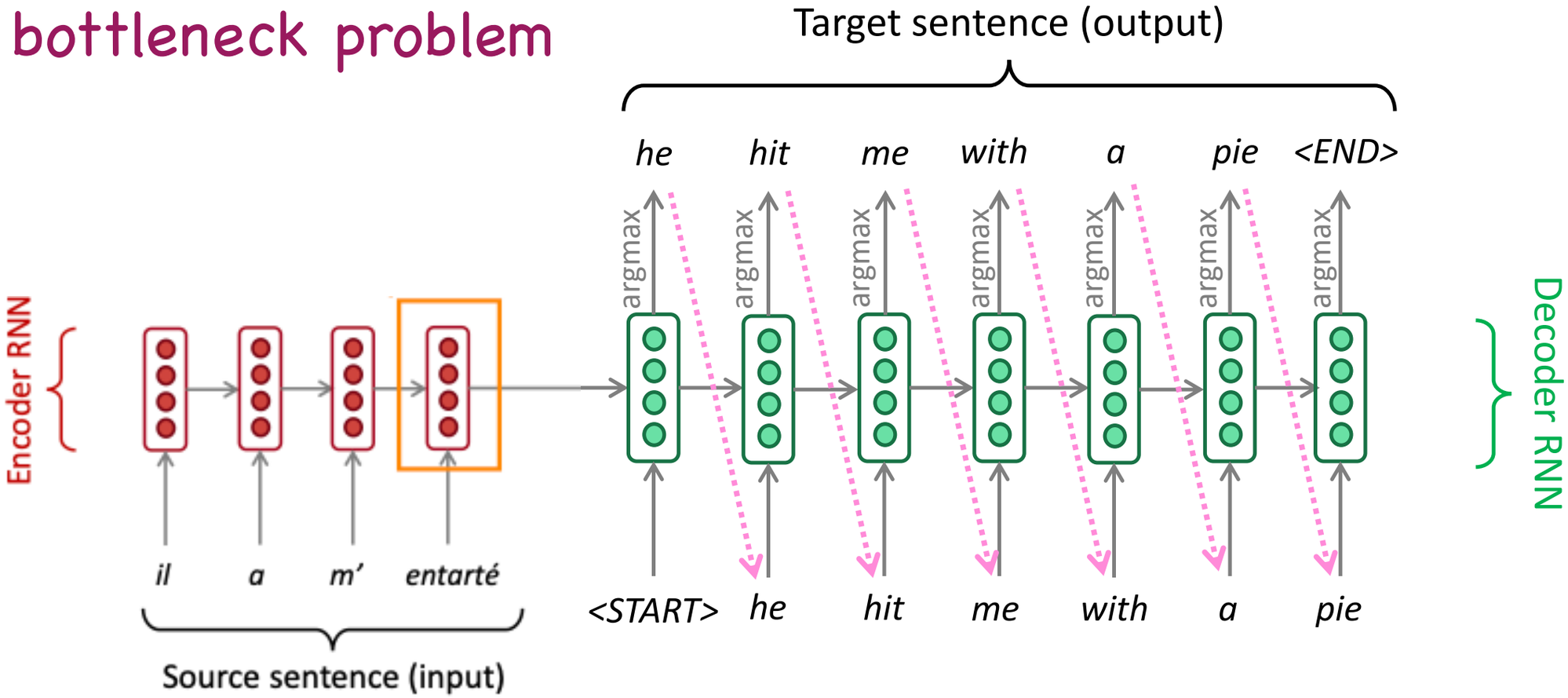
*"Improvement over the first version of AlphaFold is mostly usage of **transformer/attention mechanisms** applied to residue space and combining it with the working ideas from the first version"*

Attention/Transformer was proposed first for NMT!

https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Sequence-to-sequence Model

- The bottleneck problem



Encoding of the source sentence. This needs to capture all information about the source sentence. Information bottleneck!

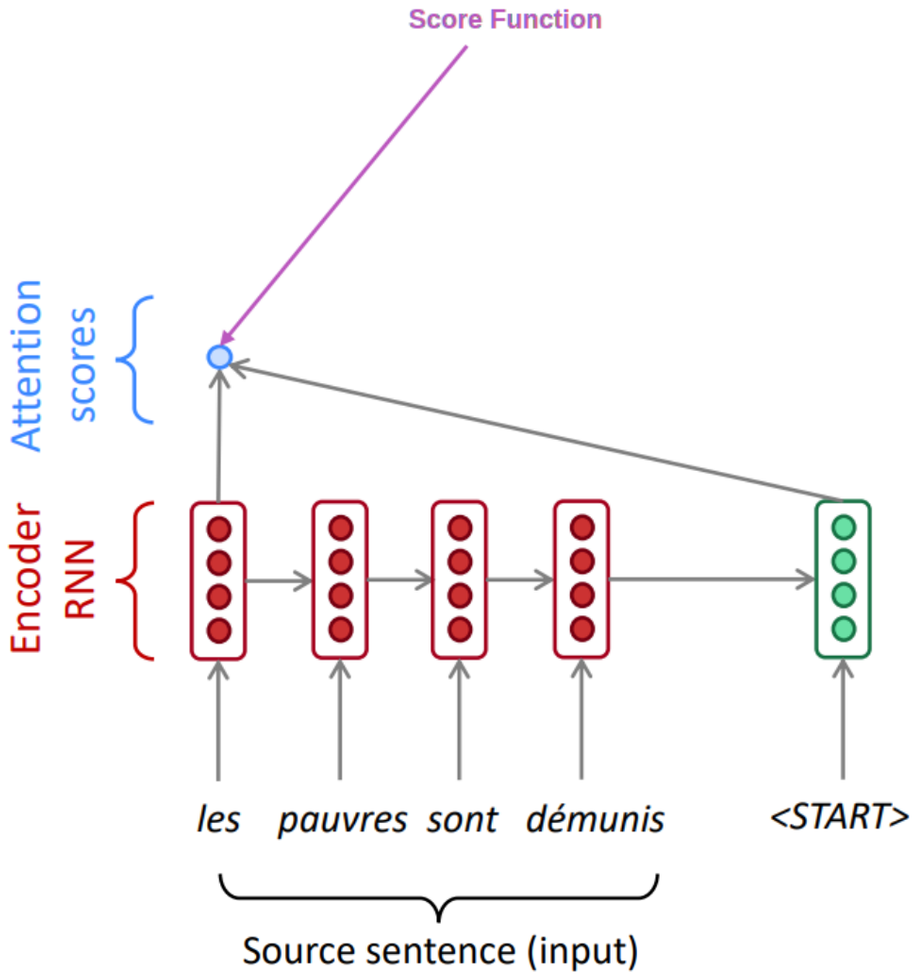
Source: Stanford 224n

Attention Mechanism

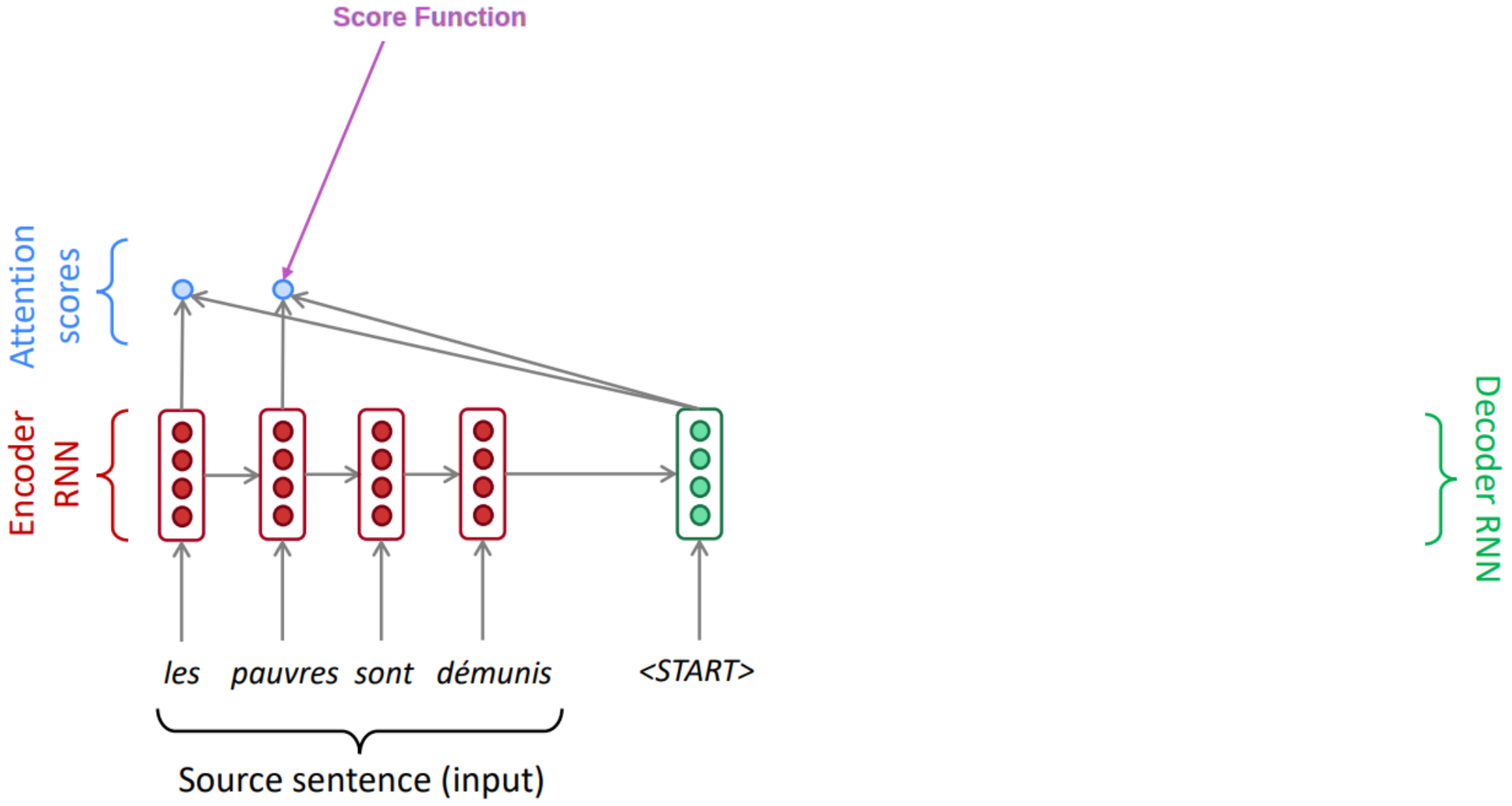
- **Attention** provides a solution to the bottleneck problem.

- **Core idea:** on each step of the decoder, use direct connection to the encoder to focus on the relevant part of the source sequence

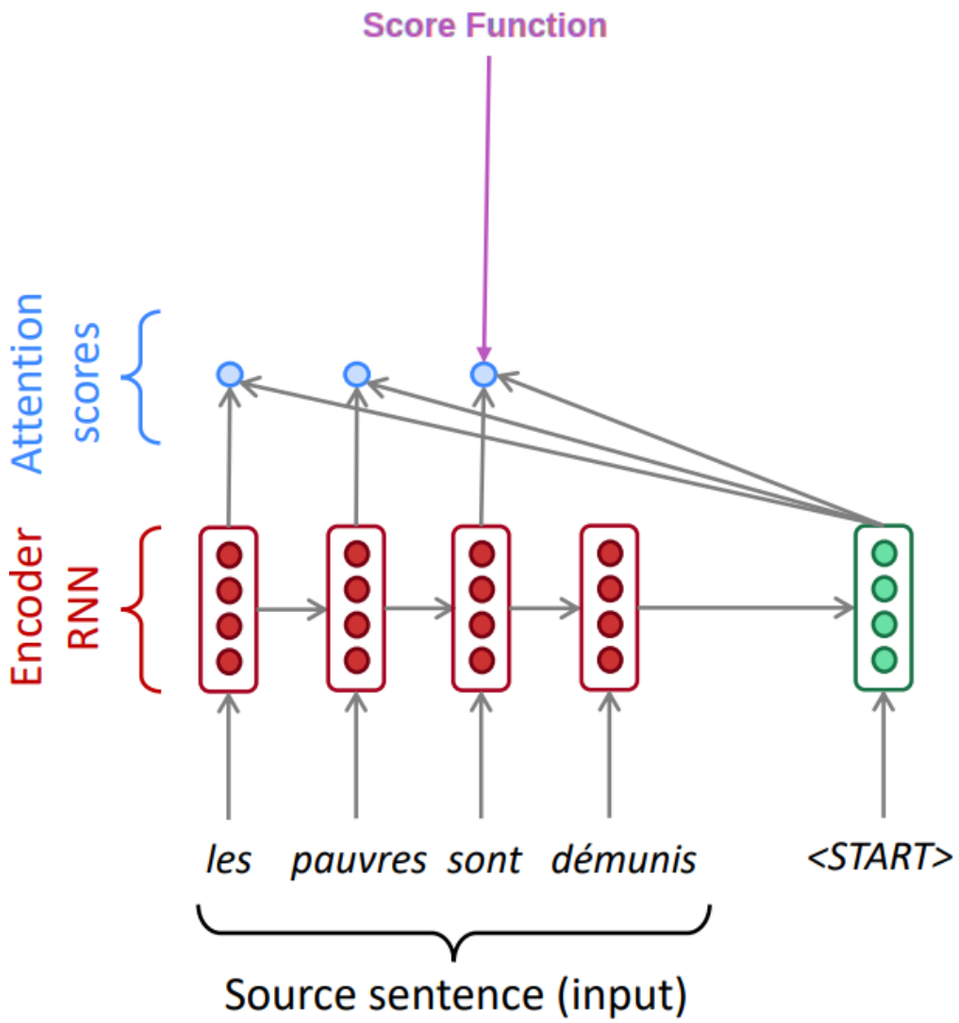
Attention Mechanism



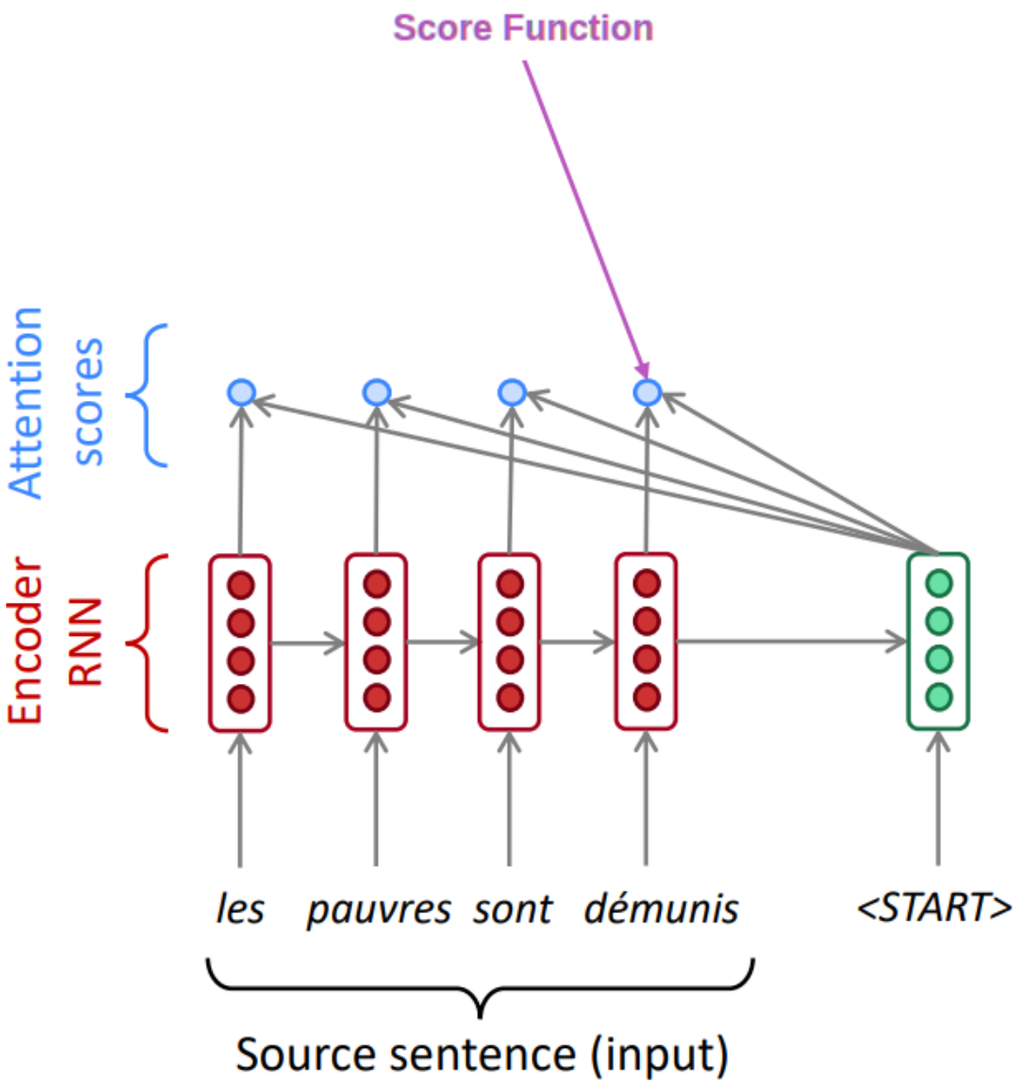
Attention Mechanism



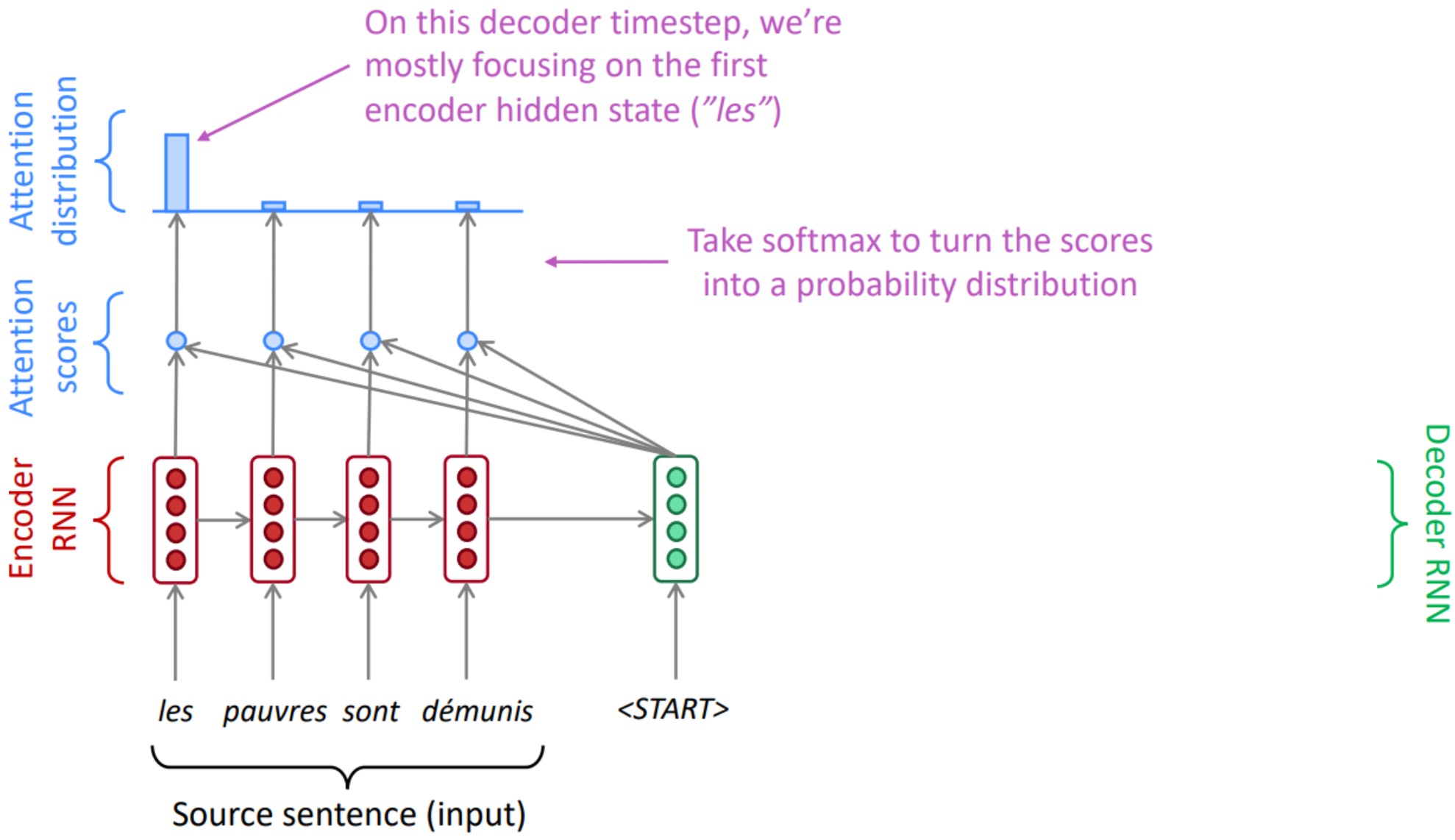
Attention Mechanism



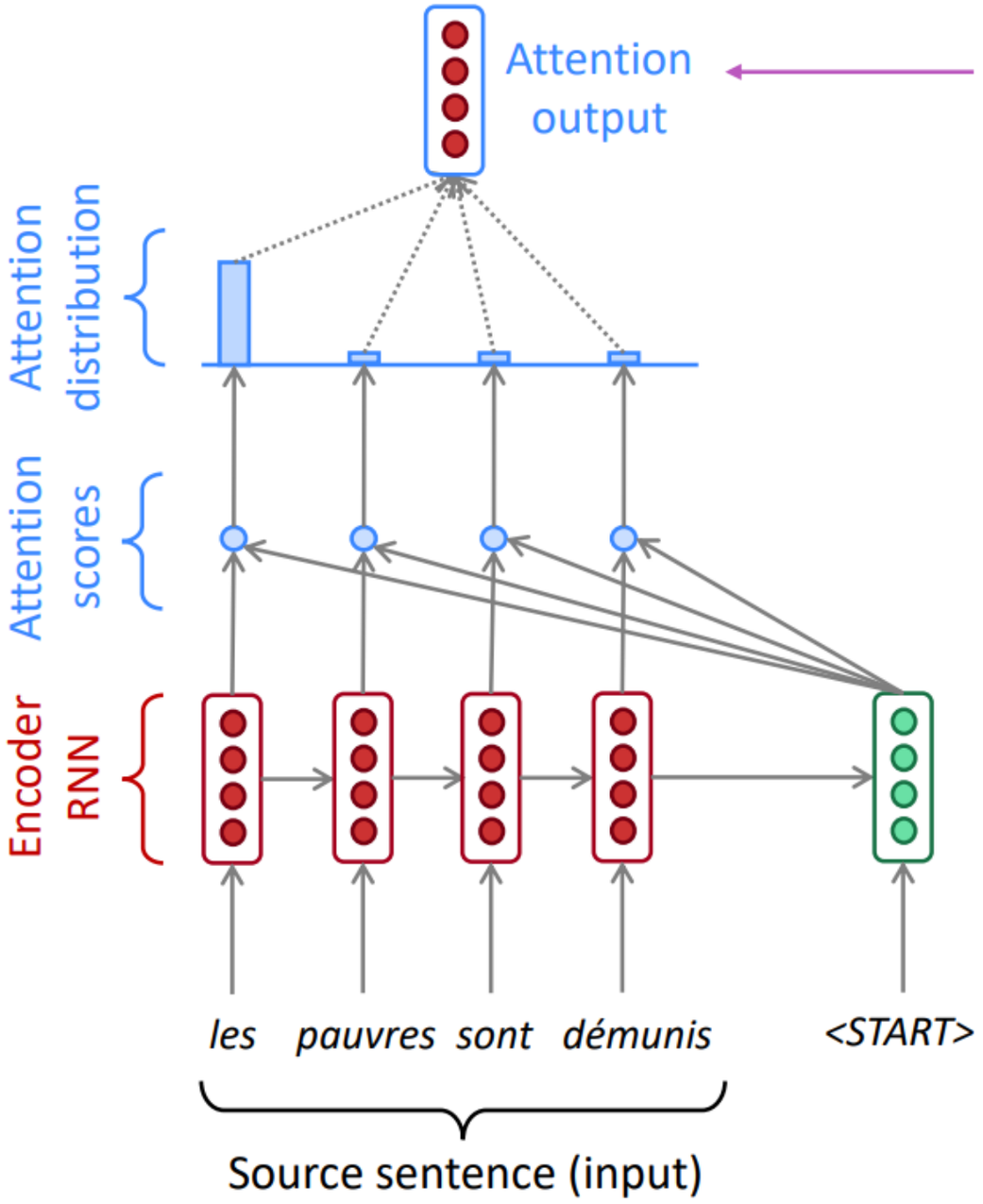
Attention Mechanism



Attention Mechanism



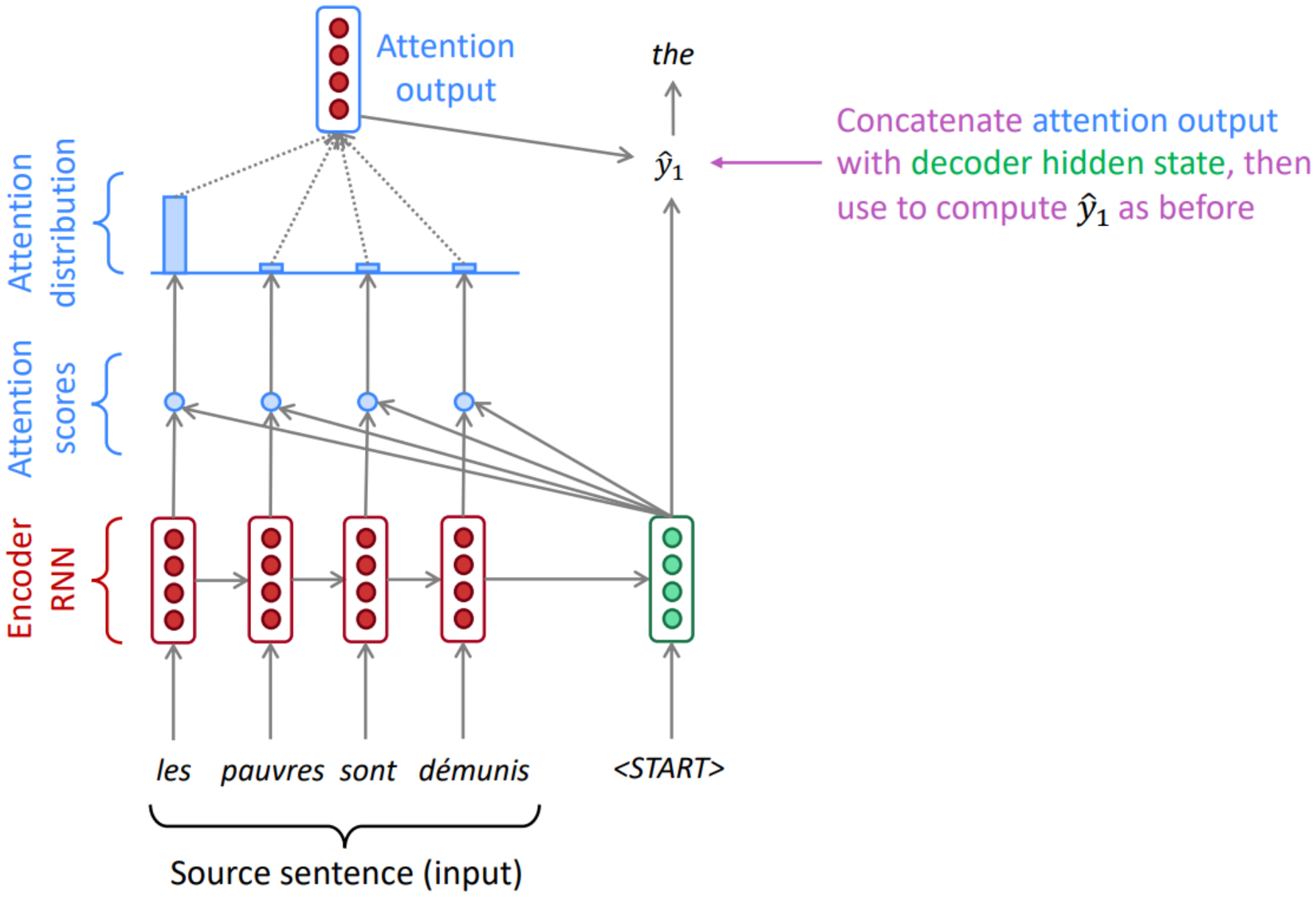
Attention Mechanism



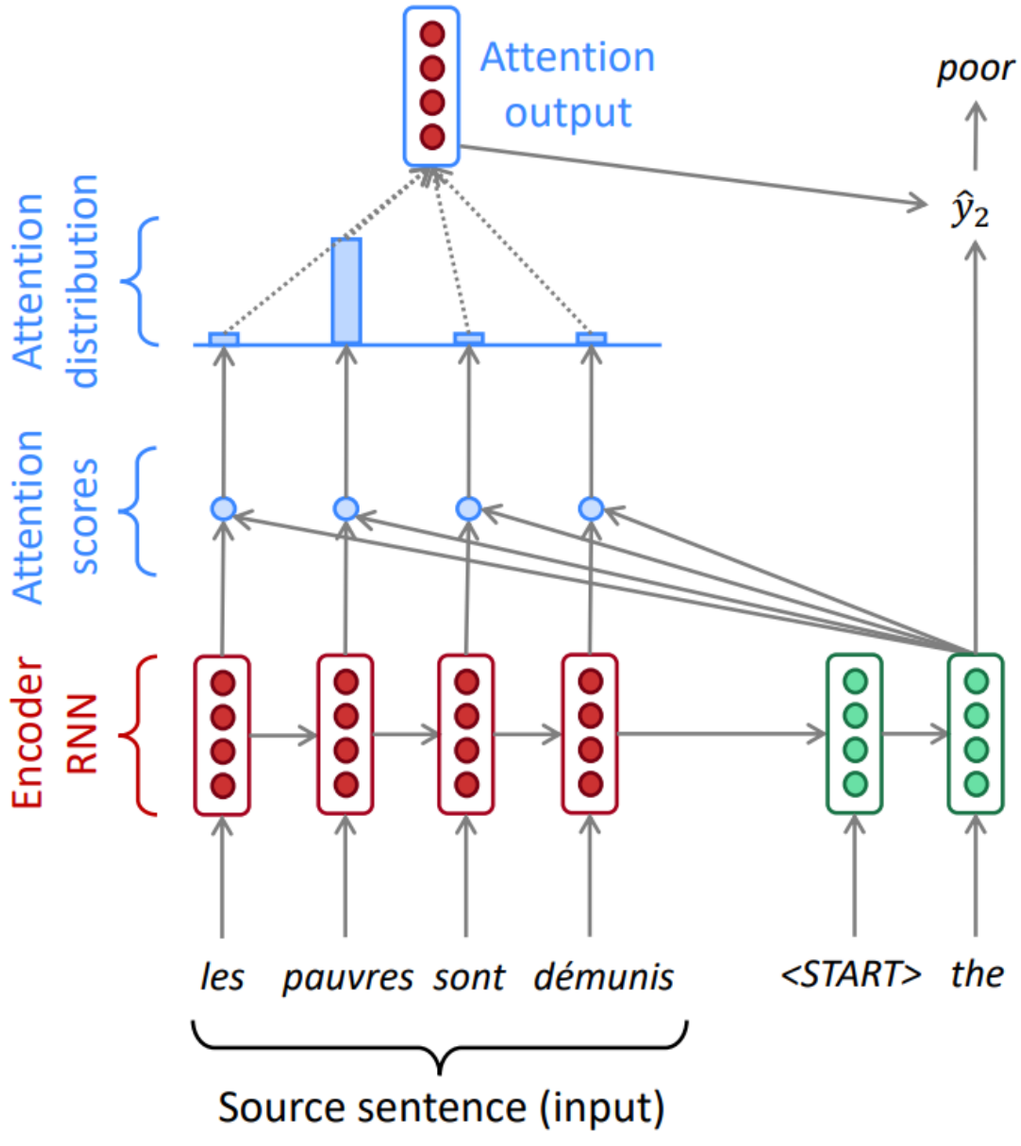
Use the attention distribution to take a **weighted sum** of the encoder hidden states.

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

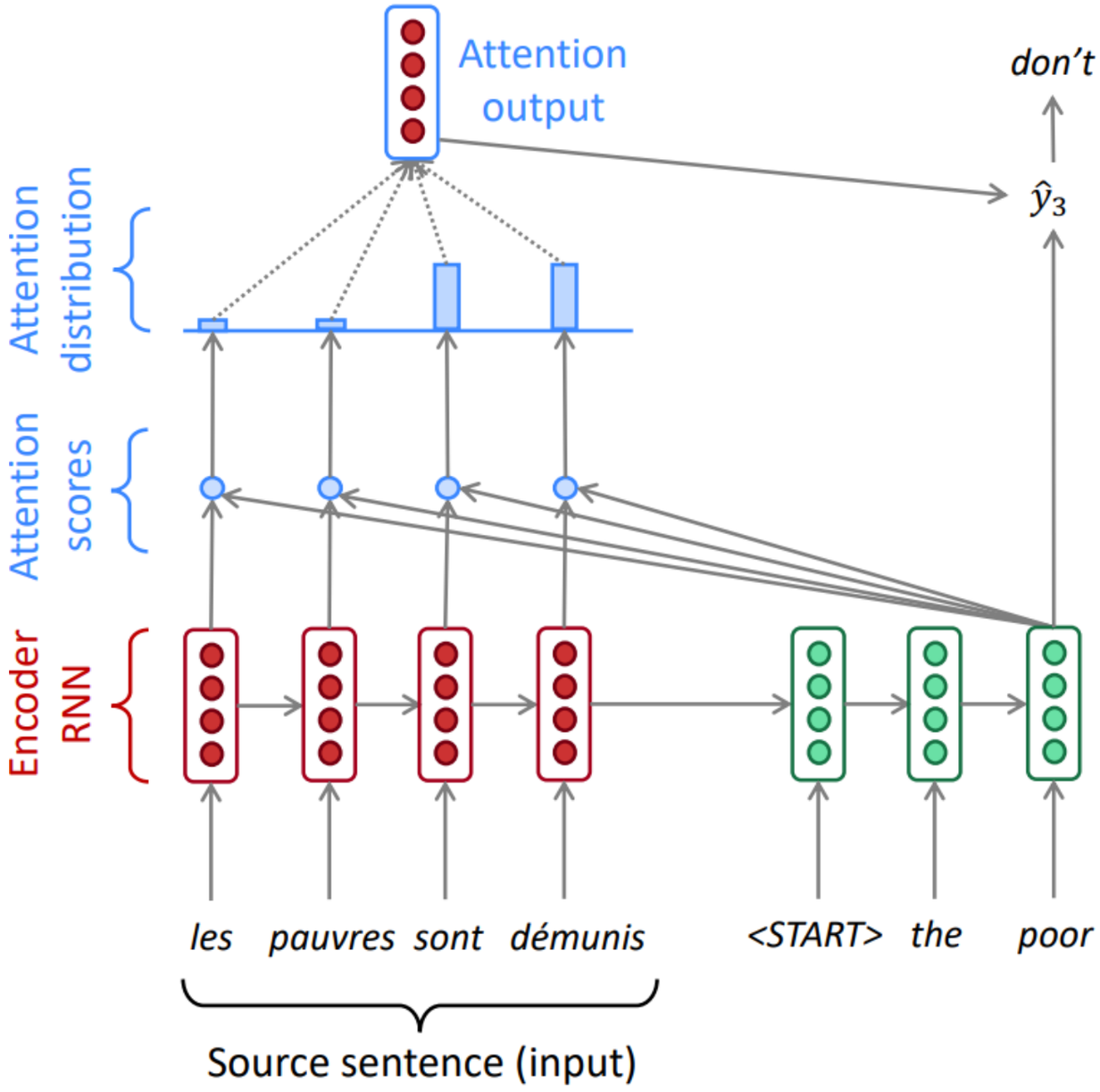
Attention Mechanism



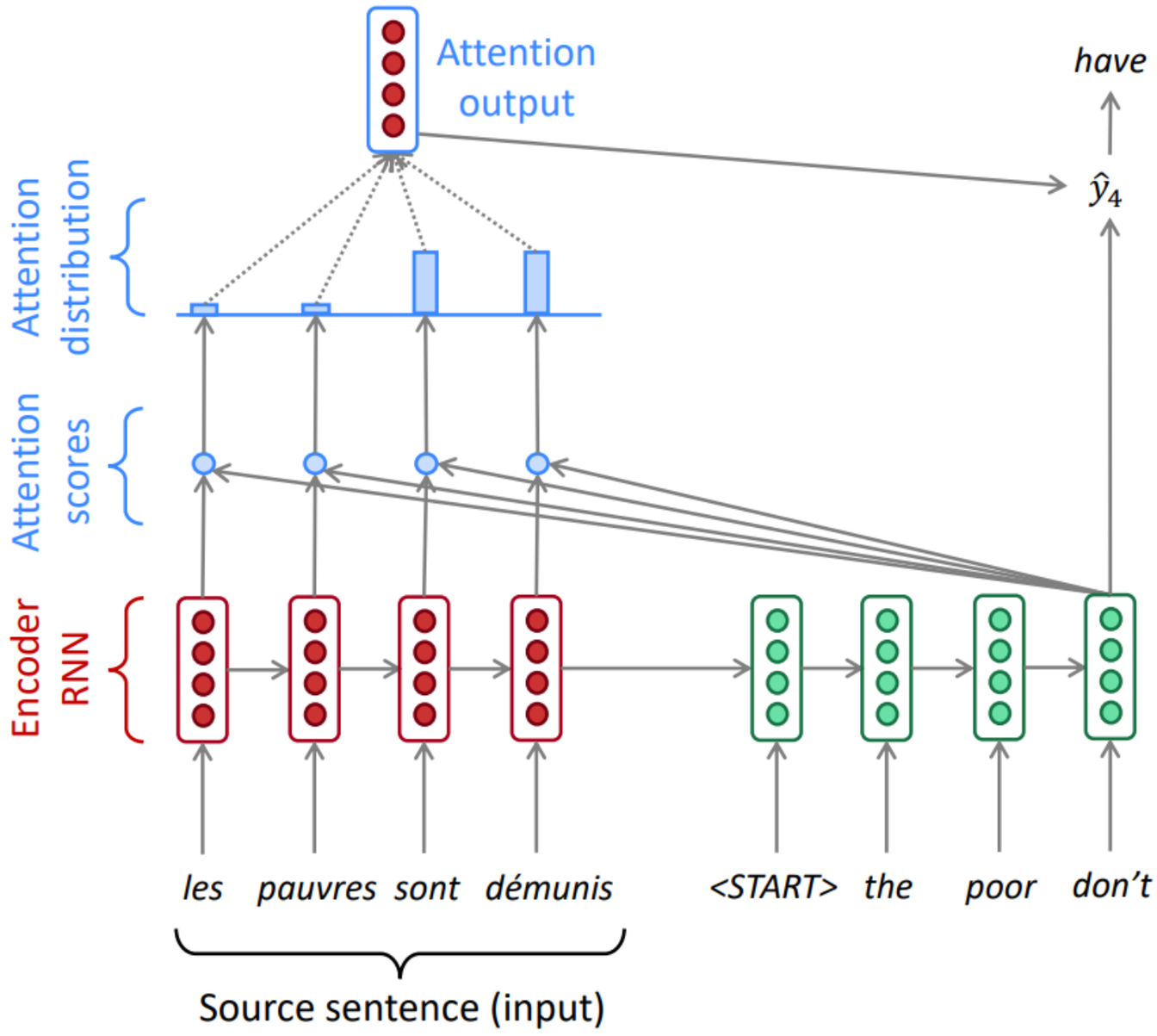
Attention Mechanism



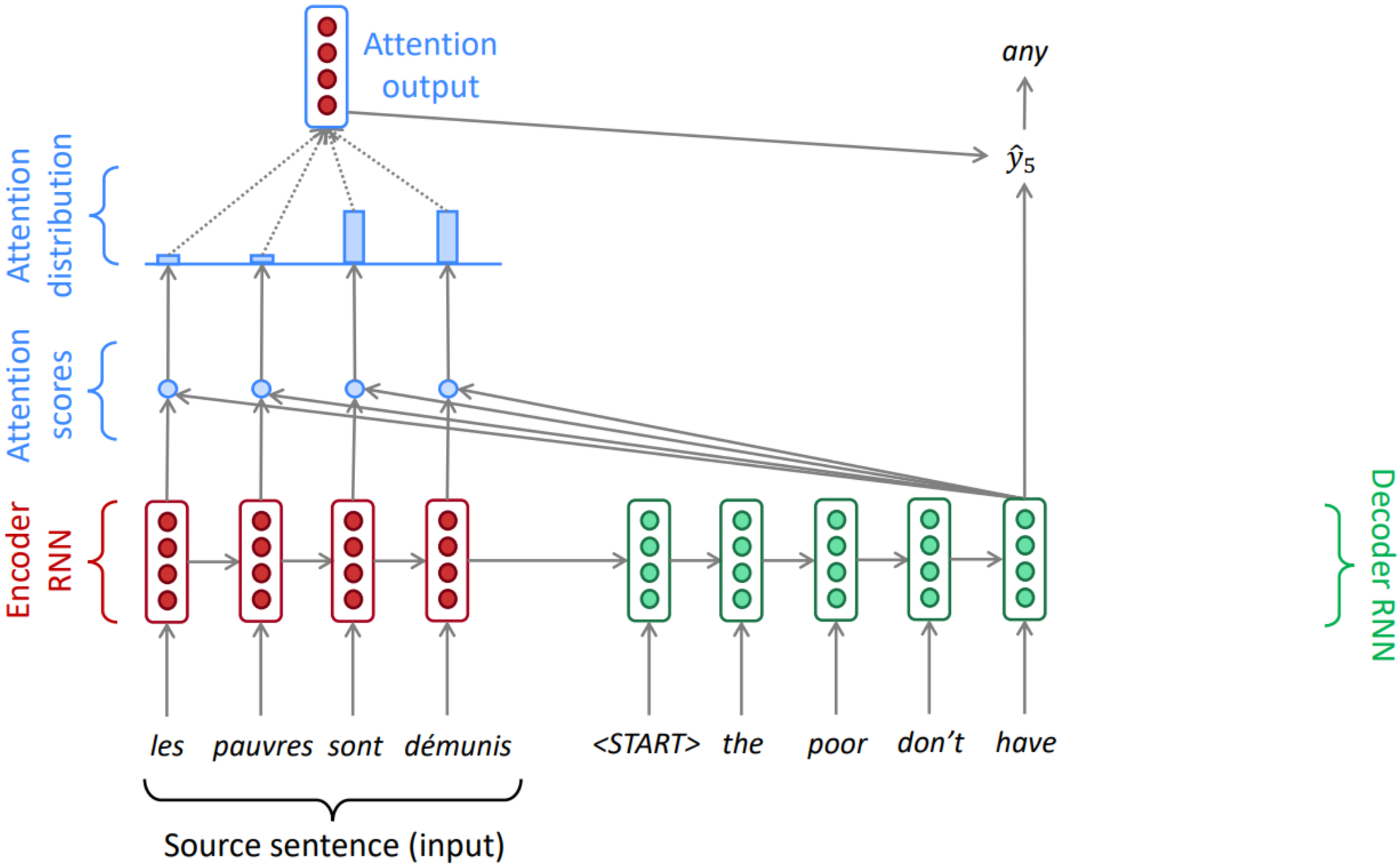
Attention Mechanism



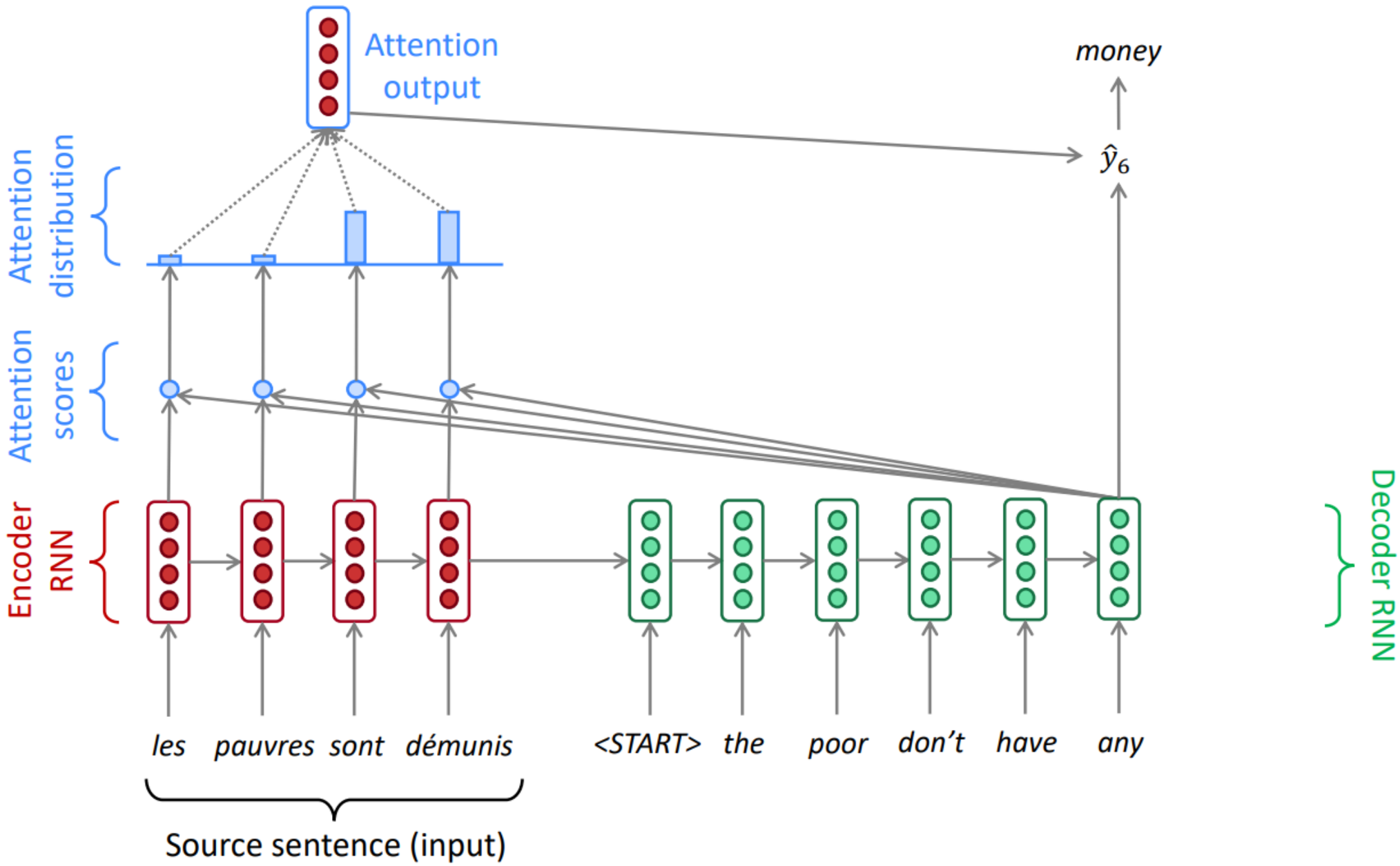
Attention Mechanism



Attention Mechanism



Attention Mechanism



Attention Mechanism (Formally)

- We have encoder hidden states $h_1, \ldots, h_N \in \mathbb{R}^h$
- On timestep *t*, we have decoder hidden state $s_t \in \mathbb{R}^h$
- We get the attention scores $\boldsymbol{e}^t$ for this step:

$$\boldsymbol{e}^t = [\boldsymbol{s}_t^T \boldsymbol{h}_1, \ldots, \boldsymbol{s}_t^T \boldsymbol{h}_N] \in \mathbb{R}^N$$

- We take softmax to get the attention distribution α^t for this step (this is a probability distribution and sums to 1)

$$\alpha^t = \text{softmax}(\boldsymbol{e}^t) \in \mathbb{R}^N$$

- We use α^t to take a weighted sum of the encoder hidden states to get the attention output $\boldsymbol{a}_t$

$$\boldsymbol{a}_t = \sum_{i=1}^{N} \alpha_i^t \boldsymbol{h}_i \in \mathbb{R}^h$$

Check Tutorial 4

Attention (Formally)

- We have encoder hidden states $h_1, \ldots, h_N \in \mathbb{R}^h$ **Keys/Values**
- On timestep *t*, we have decoder hidden state $s_t \in \mathbb{R}^h$ **Query**
- We get the attention scores e^t for this step:

$$\boldsymbol{e}^t = [\boldsymbol{s}_t^T \boldsymbol{h}_1, \ldots, \boldsymbol{s}_t^T \boldsymbol{h}_N] \in \mathbb{R}^N$$

There are several ways to do this

- We take softmax to get the attention distribution α^t for this step (this is a probability distribution and sums to 1)

$$\alpha^t = \text{softmax}(\boldsymbol{e}^t) \in \mathbb{R}^N$$

- We use α^t to take a weighted sum of the encoder hidden states to get the attention output $\boldsymbol{a}_t$

$$\boldsymbol{a}_t = \sum_{i=1}^{N} \alpha_i^t \boldsymbol{h}_i \in \mathbb{R}^h$$

Attention Variants

- We have encoder hidden states $h_1, \ldots, h_N \in \mathbb{R}^h$
- On timestep *t*, we have decoder hidden state $s_t \in \mathbb{R}^h$

- Basic dot-product attention: $\boldsymbol{e}_i = \boldsymbol{s}^T \boldsymbol{h}_i \in \mathbb{R}$
  - Note: this assumes $d_1 = d_2$
  - This is the version we saw earlier

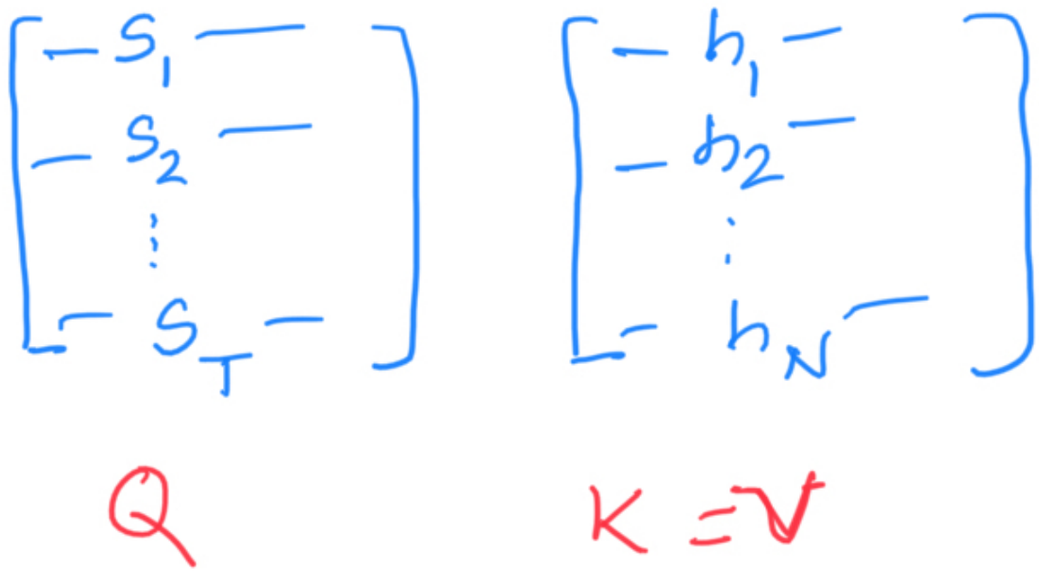
- Multiplicative attention: $\boldsymbol{e}_i = \boldsymbol{s}^T \boldsymbol{W} \boldsymbol{h}_i \in \mathbb{R}$
  - Where $\boldsymbol{W} \in \mathbb{R}^{d_2 \times d_1}$ is a weight matrix

- Additive attention: $\boldsymbol{e}_i = \boldsymbol{v}^T \tanh(\boldsymbol{W}_1 \boldsymbol{h}_i + \boldsymbol{W}_2 \boldsymbol{s}) \in \mathbb{R}$
  - Where $\boldsymbol{W}_1 \in \mathbb{R}^{d_3 \times d_1}, \boldsymbol{W}_2 \in \mathbb{R}^{d_3 \times d_2}$ are weight matrices and $\boldsymbol{v} \in \mathbb{R}^{d_3}$ is a weight vector.
  - d_3 (the attention dimensionality) is a hyperparameter

Attention in Matrix Notation

- We have encoder hidden states $h_1, \ldots, h_N \in \mathbb{R}^h$
- On timestep *t*, we have decoder hidden state $s_t \in \mathbb{R}^h$

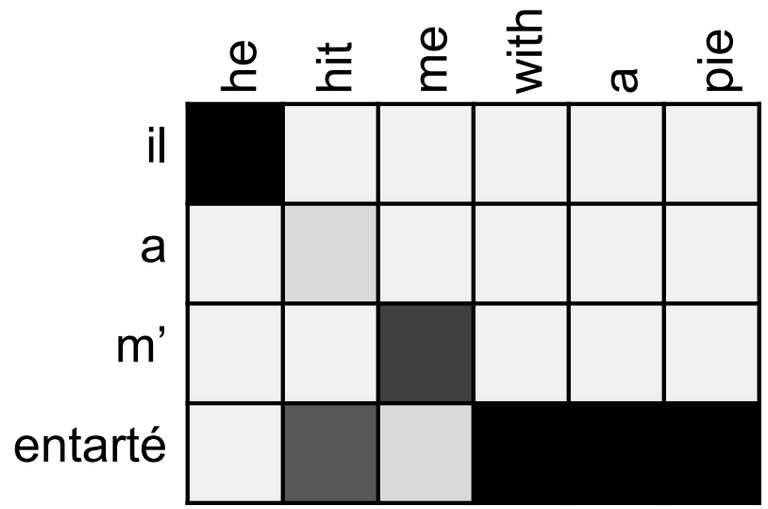
- Dot-product attention in matrix notation



$$A = S(QK^T)V$$

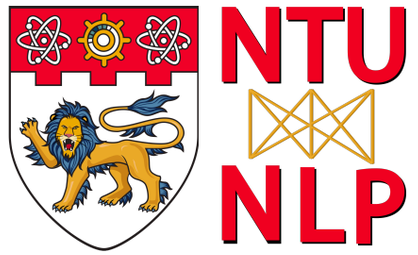
Attention turns out to be very important!

- Attention provides interpretability
  - By inspecting attention distribution, we can see what the decoder was focusing on
  - (soft) alignment for free
- Although, it has been questioned recently [1,2,3]



[1] Attention is not Explanation. Sarthak Jain, Byron C. Wallace. In NAACL-2019
[2] Attention is not not Explanation. Sarah Wiegreffe, Yuval Pinter. In EMNLP-2019
[3] Learning to Deceive with Attention-Based Explanations. Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, Zachary C. Lipton. In ACL-2020

Attention is a general Deep Learning technique



- We have encoder hidden states $h_1, \ldots, h_N \in \mathbb{R}^h$ **Keys/Values**
- On timestep *t*, we have decoder hidden state $s_t \in \mathbb{R}^h$ **Query**

- Given a set of key-value vectors, and a query vector, **attention** is a technique to compute a weighted sum of the value vectors, dependent on the query-key relevance.

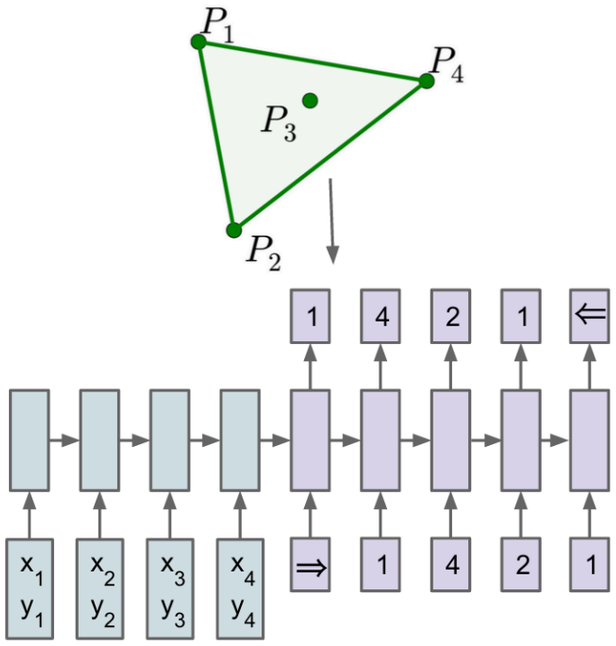
Intuition:

- The weighted sum is a selective summary of the information contained in the values, where the query-key determines which values to focus on.

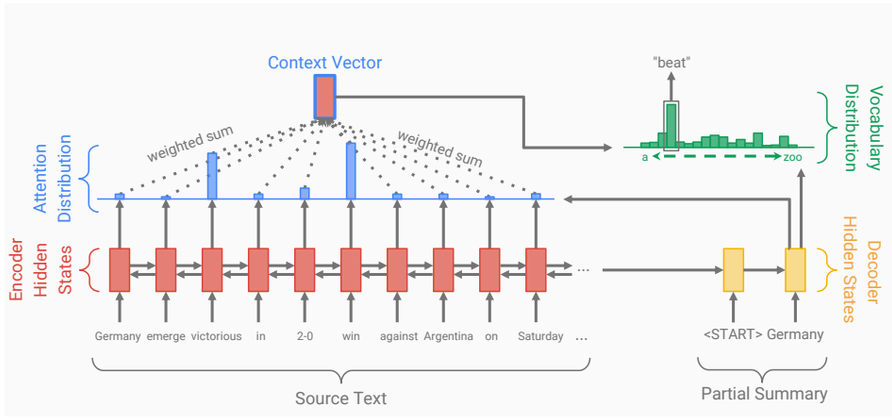
- Attention is a way to obtain a fixed-size representation of an arbitrary set of representations (values), dependent on some other representation (query-key).

Attention turns out to be very important!

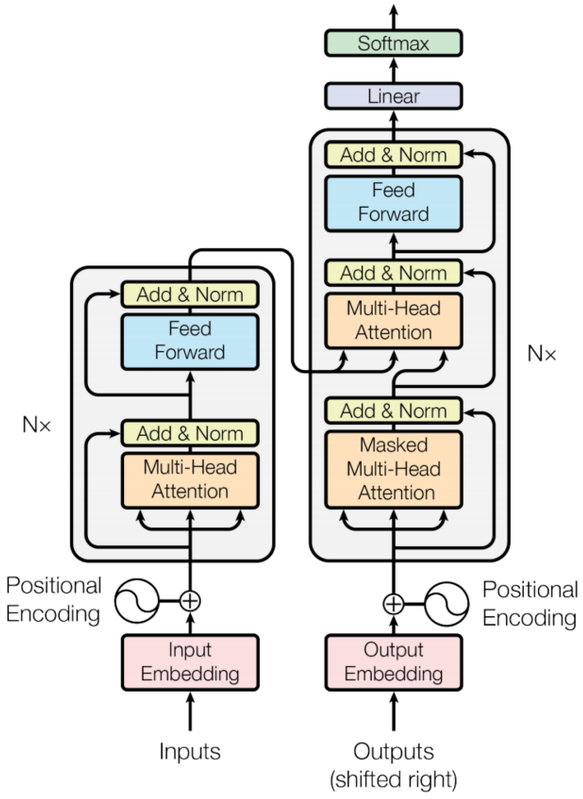
- Attention is a general Deep Learning technique
  - Not just in a seq2seq model
  - Not just in NMT; now almost everywhere in DL
  - Can be repurposed to point, to copy, or as a representation layer.



Pointer Net

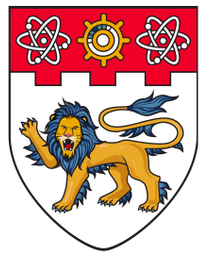


Pointer-Generator Net



Transformers

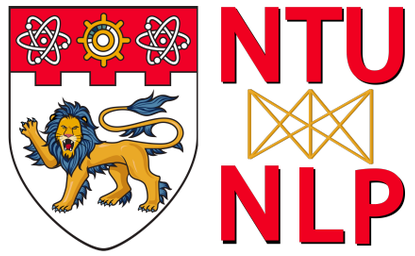
Rest of the Lecture



NTU NLP

- Seq2Seq variants
  - Pointer nets
  - Pointer generator nets
  - Stack pointer nets
- Transformer Architecture
  - Self-attention
  - Positional encoding
  - Multi-head attentions

- Applications
  - Machine Translation
  - Summarization
  - Parsing
  - Dialogue systems



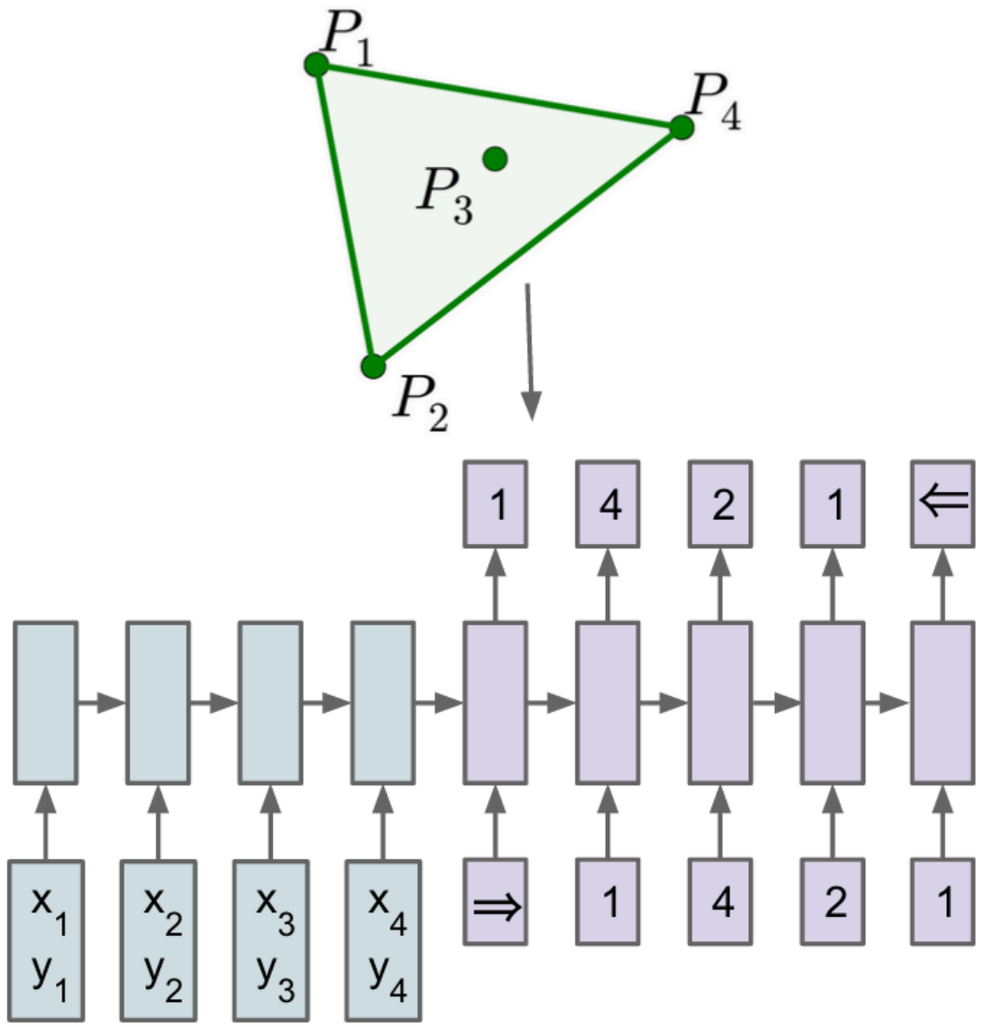
Seq2Seq Variants + Applications

NTU NLP

Pointer Nets

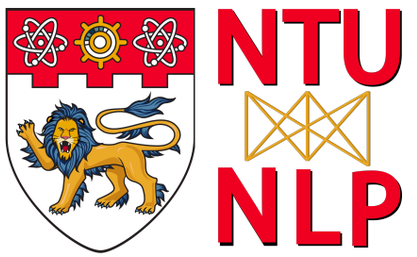
Limitation of Seq2Seq

- Fixed output dictionary
- Cannot directly apply it to combinatorial problems where the size of the output dictionary depends on the length of the input sequence

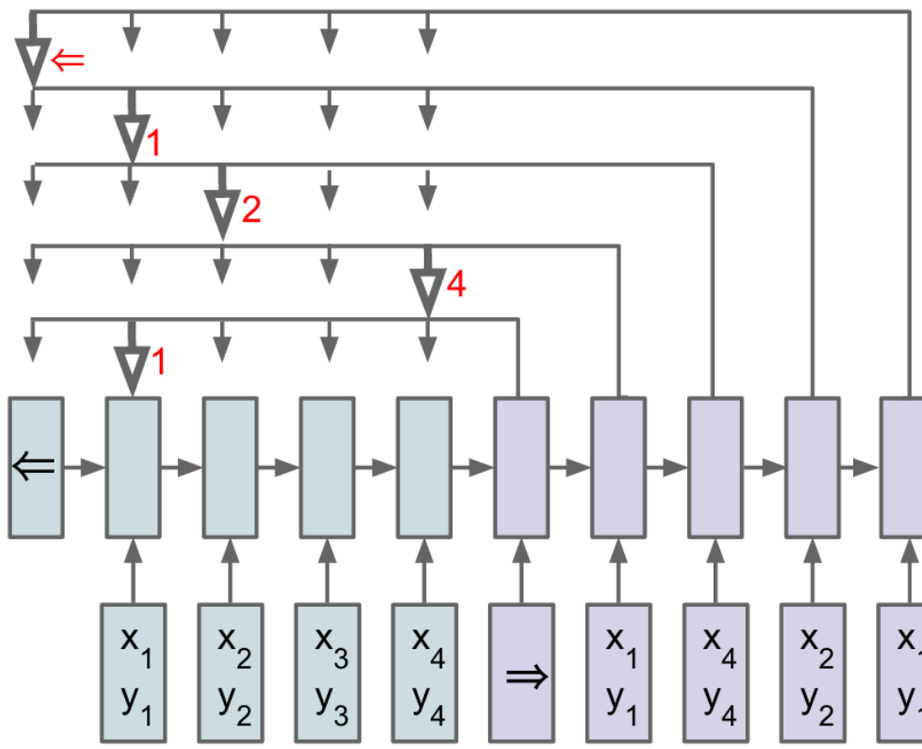


Source: Vinyals et al.

Pointer Nets



Output sequence is made of discrete tokens corresponding to positions in an input sequence



- Use attention as a pointer to select a member of the input sequence as the output
- Reduction of the seq2seq+attention model

$$e_i^t = score(h_t, s_i)$$

$$a^t = softmax(e^t)$$

NTU NLP

Seq2Seq for Summarization

Seq2Seq for Summarization

Two Approaches:

Extractive Summarization

Select parts (typically sentences) of the original text to form a summary.

- Easier
- Too restrictive (no paraphrasing)
- Most past work is extractive

Abstractive Summarization

Generate novel sentences using natural language generation techniques.

- More difficult
- More flexible and human
- Necessary for future progress

Source: See et al (2017)

Seq2Seq for Summarization

NTU NLP

CNN / Daily Mail dataset

- Long news articles (average ~800 words)
- Multi-sentence summaries (usually 3 or 4 sentences, average 56 words)
- Summary contains information from throughout the article

Daily Mail

The papal bill Pope Francis insisted on paying himself... before catching the bus home after winning the election

- **Pope Francis insisted on returning to his hotel to settle the bill himself**
- **The pontiff also chose to use a bus instead of a chauffeur driven car**
- **The 76-year-old has eschewed ceremonial traditions for a more humble approach**

With the spiritual wellbeing of the world's 1.2 billion Catholics is on his shoulders he must have quite a to-do list.

But despite his new responsibilities, Francis did not forget to stop off - between engagements - to pay his hotel bill.

Staff at the central Rome priests' residence where Bergoglio was staying before the conclave, were astonished when the newly elected Pope strolled in to collect his luggage and settle the bill.

Pope Francis insisted on returning to the hotel to collect his luggage and greet the staff before settling the hotel bill himself

'I need to set a good example' he joked.

He was driven to the hotel in a simple car and The Rev. Pawel Rytel-Andrianek, who teaches at the nearby Pontifical Holy Cross University and is staying at the residence, said that workers at the hotel were touched by the Pope's decision to return and bid them farewell.

'He wanted to come here because he wanted to thank the personnel, people who work in this house,' he said. 'He greeted them one by one, no rush, the whole staff, one by one.' Mr Rytel-Andrianek added that Francis apparently knew everyone by name.

A Vatican spokesman said: 'He wanted to get his luggage and the bags. He had left everything there.

'He then stopped in the office, greeted everyone and decided to pay the bill for the room... because he was concerned about giving a good example of what priests and bishops should do.'

Francis is already winning plaudits for his down-to-earth manner.

He has so far refused a motorcade and the official papal Jag for official business. And even on the night of the election he insisted on accompanying the other cardinals back to their lodgings by mini bus, saying: 'I came on the bus, so I'll go home on the bus.'

Meeting cardinals yesterday on his second day of Papal business he eschewed protocol in favour of kissing on two cheeks, shaking hands and hugging.

He told his deputies that old people like himself are 'like good wine, getting better with age', before urging them to impart their wisdom to the young.

Francis began his reign in unorthodox fashion as he shunned public events in order to pray to the Virgin Mary.

During his first Mass since being elected as supreme pontiff, Pope Francis and his cardinals were dressed in simple yellow robes over their cassocks, rather than the formal ceremonial outfits they would normally wear on such a major occasion.

Speaking in Italian without notes, he said: 'We can walk all we want, we can build many things, but if we don't proclaim Jesus Christ, something is wrong. We would become a compassionate NGO and not a Church which is the bride of Christ.

'He who does not pray to the Lord prays to the devil. When we don't proclaim Jesus Christ, we proclaim the worldliness of the devil, the worldliness of the demon.'

'We must always walk in the presence of the Lord, in the light of the Lord, always trying to live in an irreprehensible way,' he said in a heartfelt homily of a parish priest, loaded with biblical references and simple imagery.

'When we walk without the cross, when we build without the cross and when we proclaim Christ without the cross, we are not disciples of the Lord. We are worldly,' he said

'We may be bishops, priests, cardinals, popes, all of this, but we are not disciples of the Lord,' he said.

It was a far simpler message than the dense, three-page discourse Benedict delivered in Latin during his first Mass as pope in 2005.

The difference in style was a sign of Francis' belief that the Catholic Church needs to be at one with the people it serves and not impose its message on a society that often doesn't want to hear it, Francis' authorised biographer, Sergio Rubin, said.

Francis took the helm of the 1.2 billion-member Church at a time of strife and intrigue, with the Vatican rocked by a string of sex abuse scandals, accusations of infighting within its central government and by allegations of financial wrongdoing.

But many within the church believe he could change it for the better.

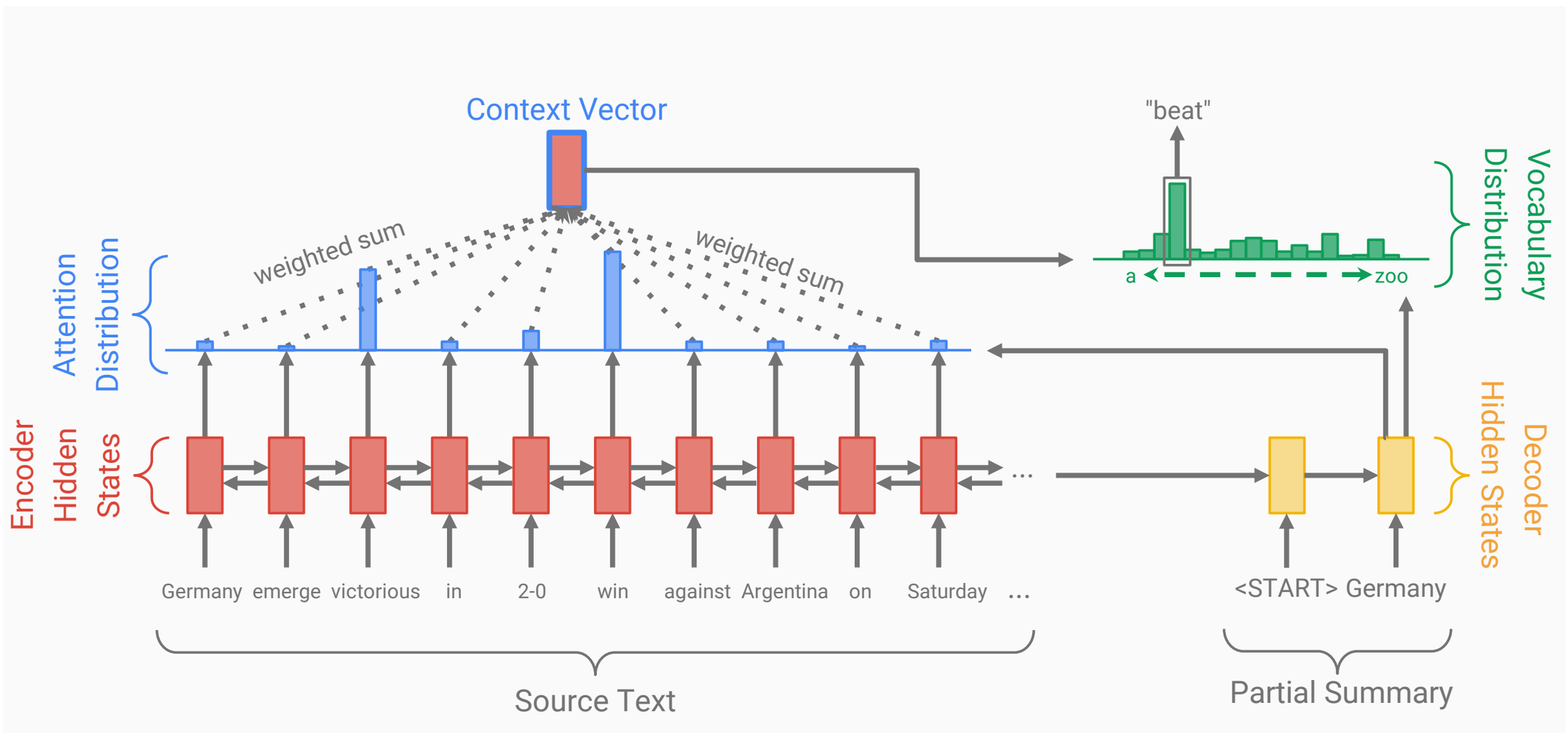
'It seems to me for now what is certain is it's a great change of style, which for us isn't a small thing,' Mr Rubin said, recalling how the former Cardinal Jorge Bergoglio would celebrate Masses with homeless people and prostitutes in Buenos Aires.

'He believes the church has to go to the streets,' he said, 'to express this closeness of the church and this accompaniment with those who are suffering.'

Source: See et al (2017)

Seq2Seq for Summarization

Seq2Seq + Attention

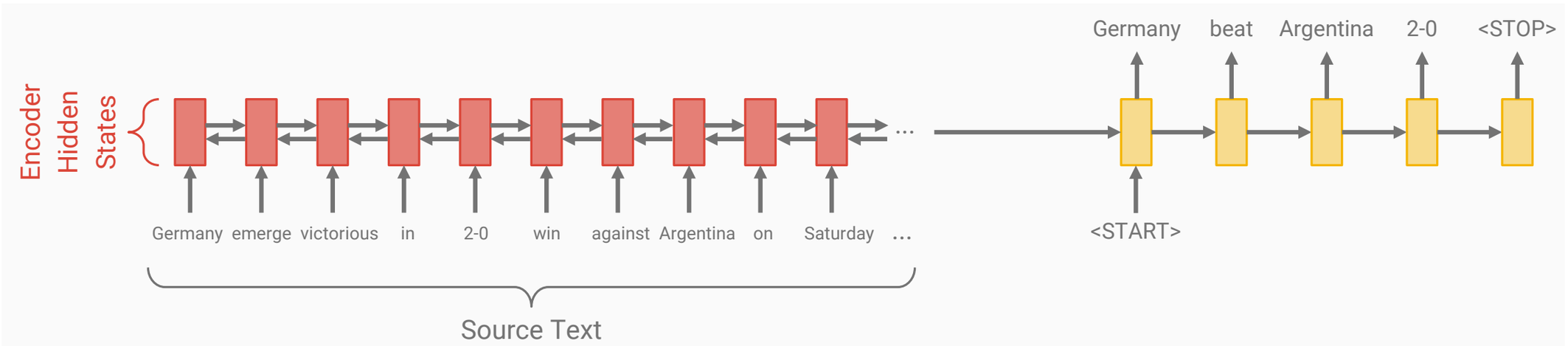


Source: See et al (2017)

Seq2Seq for Summarization

NTU NLP

Seq2Seq + Attention



Problem 1: The summaries sometimes reproduce factual details inaccurately.

e.g. *Germany beat Argentina 3-2*

Incorrect **rare** or **out-of-vocabulary** word

Solution: Use a pointer to copy words.

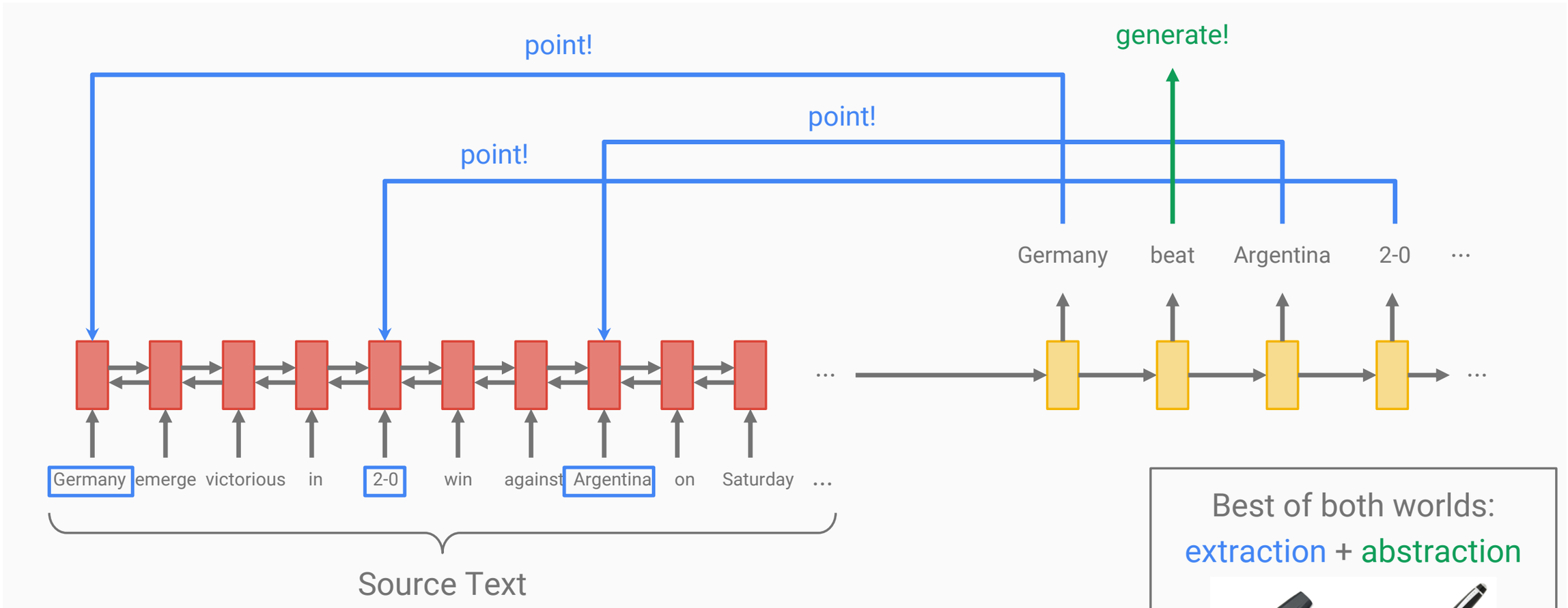
Problem 2: The summaries sometimes repeat themselves.

e.g. *Germany beat Germany beat Germany beat…*

Source: See et al (2017)

Seq2Seq for Summarization

Idea 1: Copy/Generate



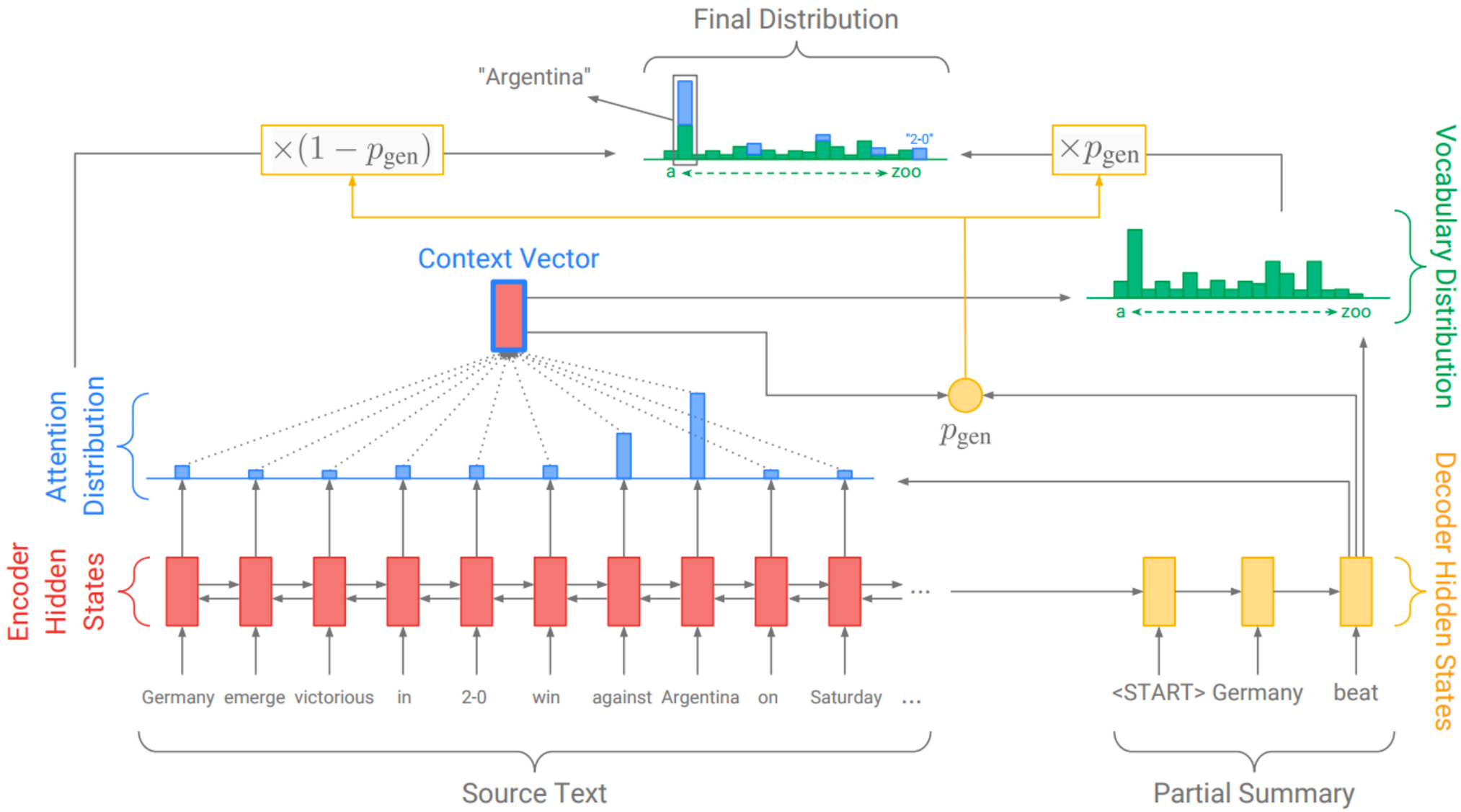
[1] ***Incorporating copying mechanism in sequence-to-sequence learning.*** *Gu et al., 2016.*
[2] ***Language as a latent variable: Discrete generative models for sentence compression.*** *Miao and Blunsom, 2016.*

Source: See et al (2017)

Seq2Seq for Summarization

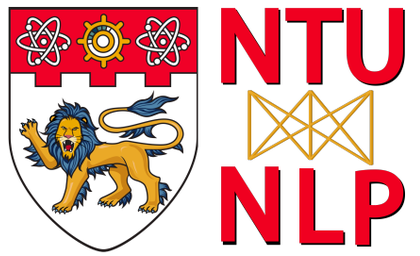
NTU NLP

Pointer Generator



Source: See et al (2017)

Seq2Seq for Summarization



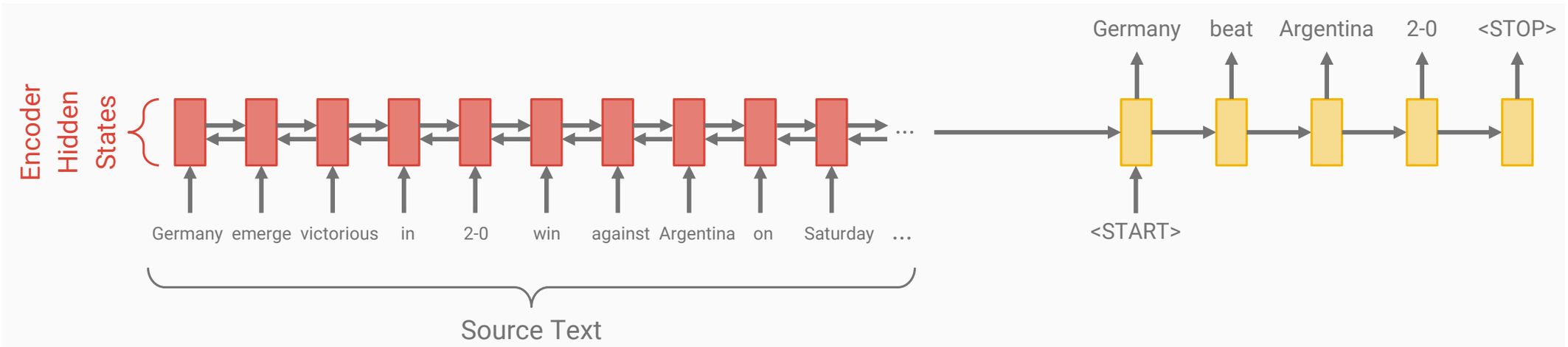
Improvements

Before	After
UNK UNK was expelled from the dubai open chess tournament	*gaioz nigalidze was expelled from the dubai open chess tournament*
the 2015 rio olympic games	*the 2016 rio olympic games*

Source: See et al (2017)

Seq2Seq for Summarization

Seq2Seq + Attention



Problem 1: The summaries sometimes reproduce factual details inaccurately.

e.g. *Germany beat Argentina 3-2*

Solution: Use a pointer to copy words.

Problem 2: The summaries sometimes repeat themselves.

e.g. *Germany beat Germany beat Germany beat…*

Solution: Penalize repeatedly attending to same parts of the source text.

Source: See et al (2017)

Seq2Seq for Summarization

NTU NLP

Reducing repetition with coverage

Coverage = cumulative attention = what has been covered so far

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday

Summary: Germany

1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

[4] ***Modeling coverage for neural machine translation.*** *Tu et al., 2016,*
[5] ***Coverage embedding models for neural machine translation.*** *Mi et al., 2016*
[6] ***Distraction-based neural networks for modeling documents.*** *Chen et al., 2016.*

Source: See et al (2017)

Seq2Seq for Summarization

NTU NLP

Reducing repetition with coverage

Coverage = cumulative attention = what has been covered so far

1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

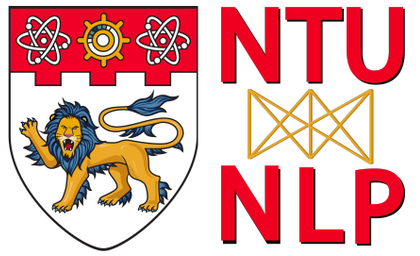
Result: repetition rate reduced to level similar to human summaries

Also see our paper:

Resurrecting Submodularity in Neural Abstractive Summarization. Simeng Han, Xiang Lin, and Shafiq Joty

Source: See et al (2017)

Seq2Seq for Summarization



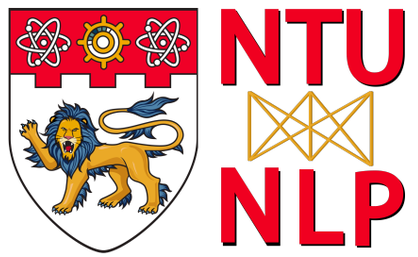
Evaluation

ROUGE compares the machine-generated system summary to the human-written reference summary and counts co-occurrence of and **1-grams**, **2-grams**, and **longest common subsequences**

Model	R-1	R-2	R-L	# Param
LEAD-3	40.00	17.50	36.28	–
Single Model				
PG†	36.69	15.92	33.63	27.9M
PG + Cov.†	39.08	17.09	35.92	27.9M + 512
6-layer Transformer¶	**40.21**	17.76	37.09	128.2M
PG + DimAttn	40.01	17.74	36.94	27.9M
PG + DyDimAttn	40.13	**17.94**	**37.21**	27.9M

Resurrecting Submodularity in Neural Abstractive Summarization. Simeng Han, Xiang Lin, and Shafiq Joty

Seq2Seq for Summarization



Evaluation

- Summarization is a subjective task
  - Humans can summarize in many ways

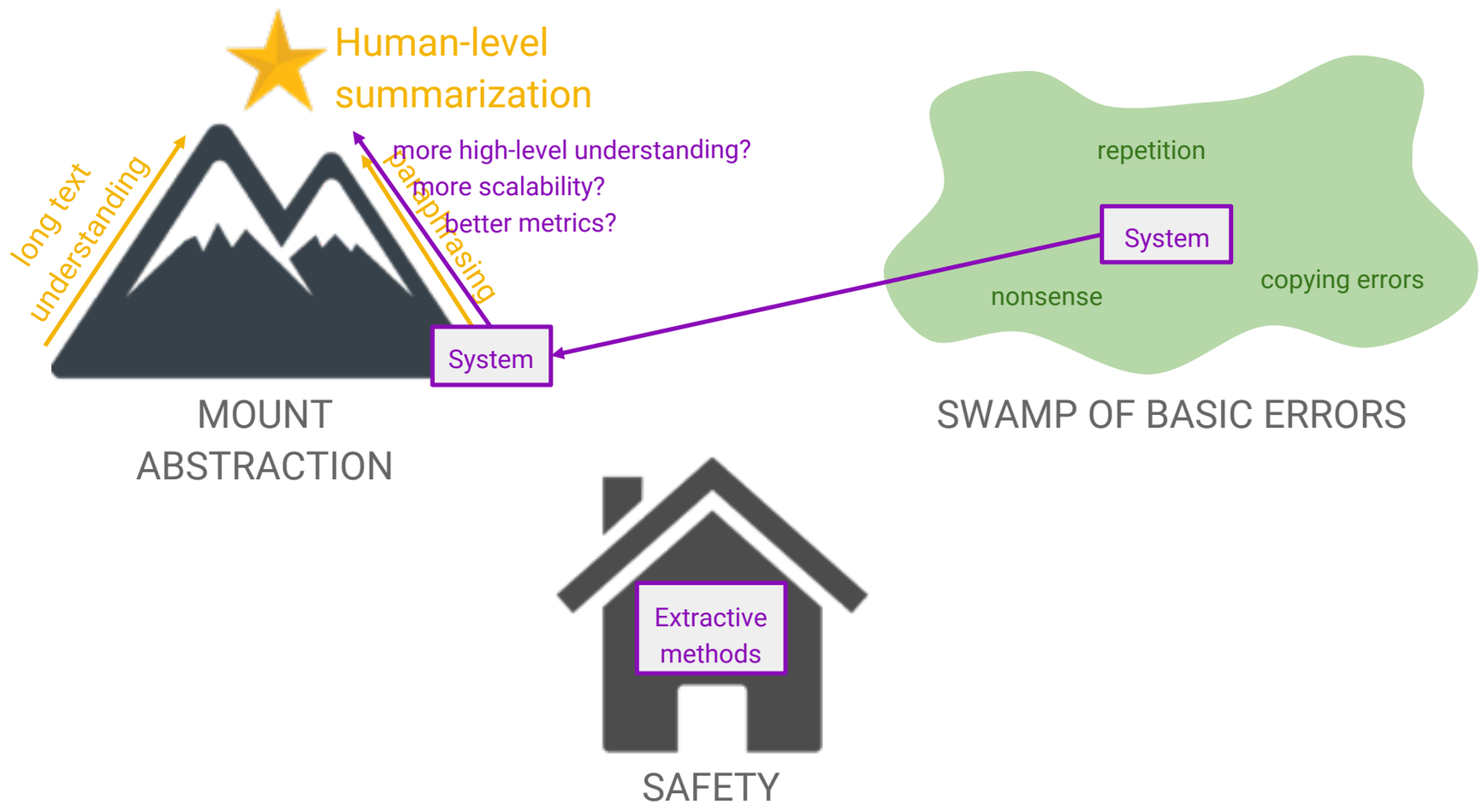
- ROUGE is based on strict comparison of n-grams to a reference summary
  - Intolerant to paraphrases
  - Favors extraction

- Lead-3 (first 3 sentences) of a news article is a strong baseline
  - Higher ROUGE score than many systems.

- Needs human study
  - Almost must for publishing at top NLP conferences

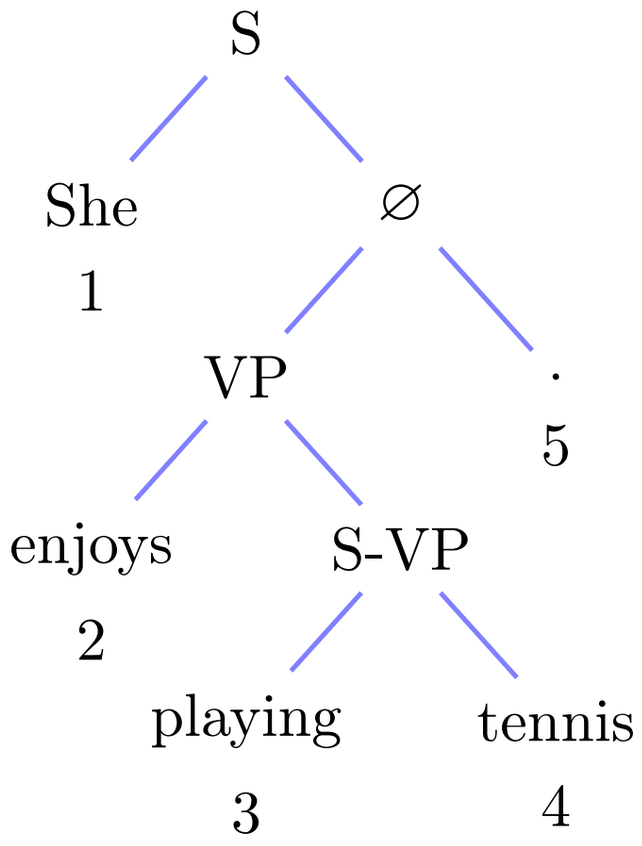
Seq2Seq for Summarization

Future Challenges

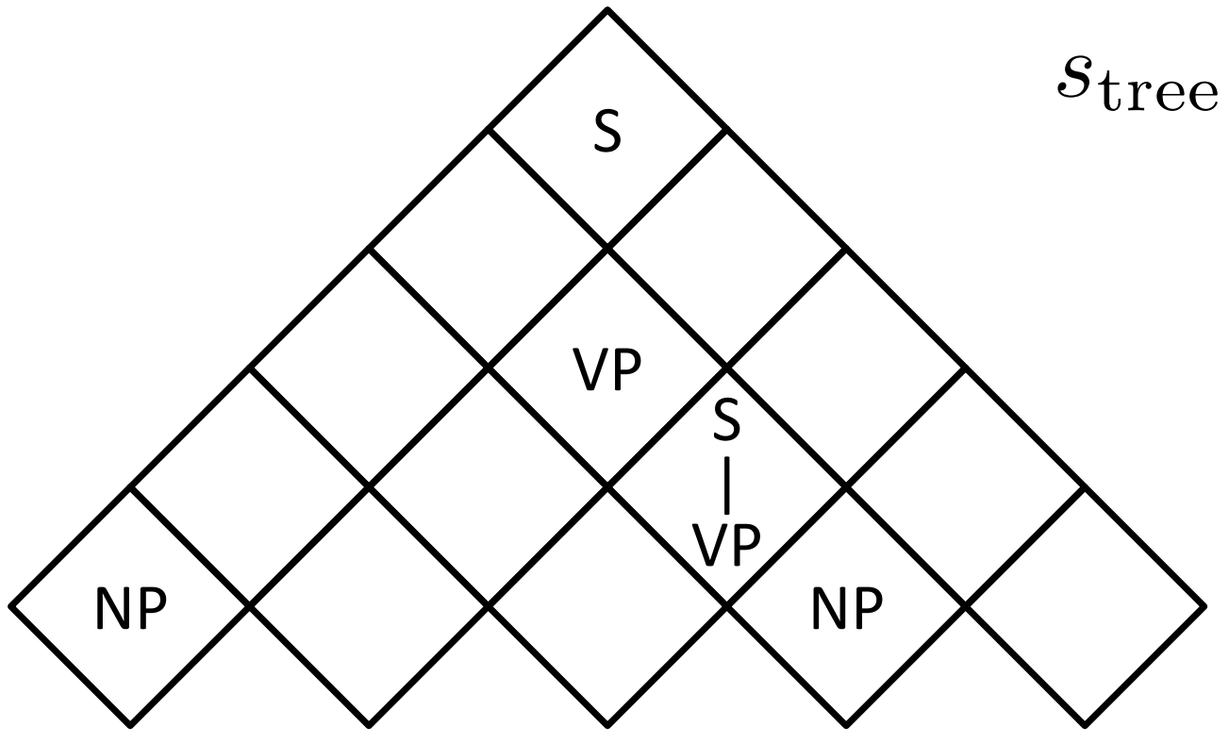


Source: See et al (2017)

Seq2Seq for Parsing



Scoring in Modern Parsers



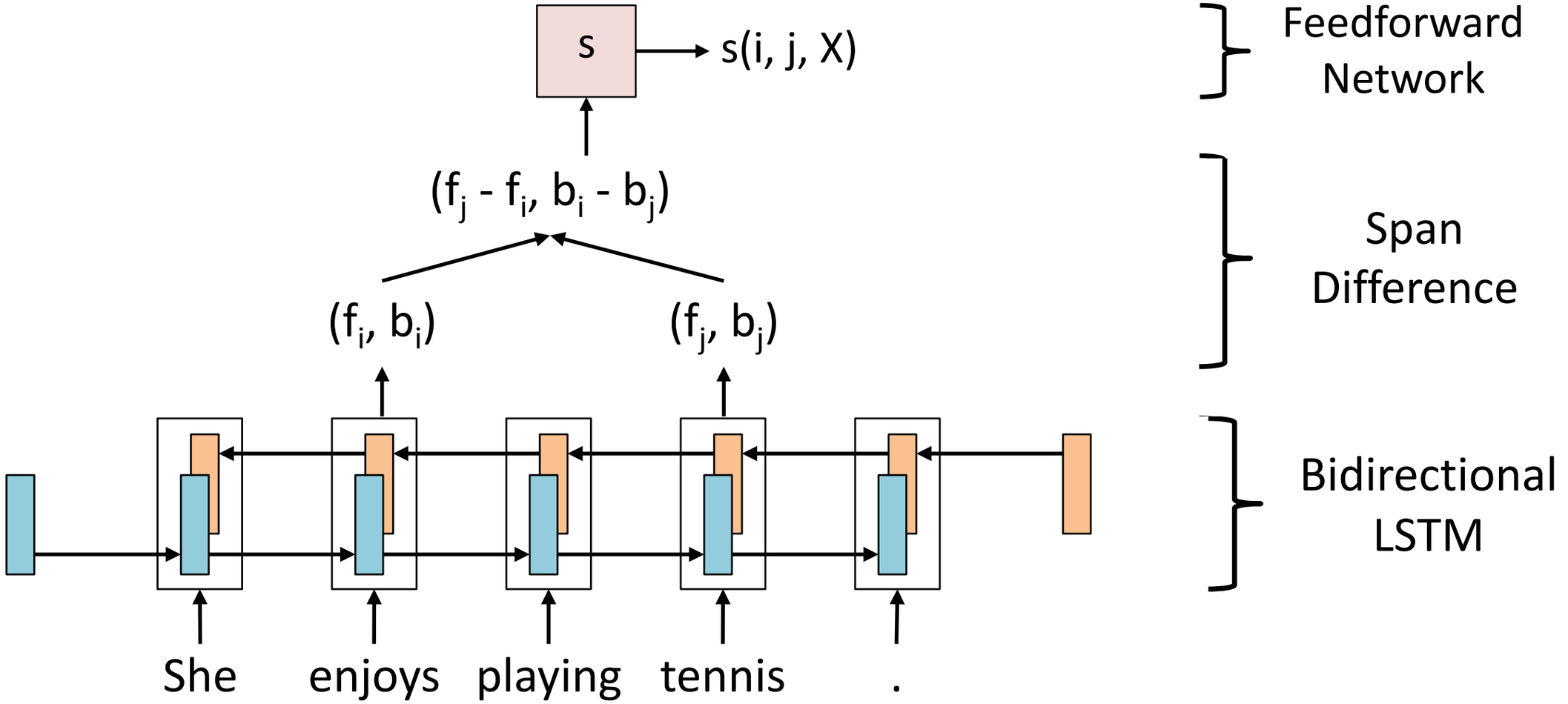
$$s_{\text{tree}}(T) = \sum_{(\ell,(i,j))\in T} s(i,j,\ell)$$

$$= s(0,5,\text{S}) + s(0,1,\text{NP}) + s(1,4,\text{VP}) + s(2,4,\text{S–VP}) + s(3,4,\text{NP})$$

Source: Stern et al, 2017

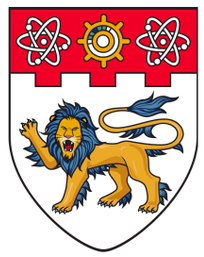
Scoring Functions

Using Sequential B-LSTM as the Sentence Encoder (Stern et al, 2017)



[Inspired by Cross and Huang (2016)]

Dynamic Programming (CKY)



NTU NLP

Bottom-up Parsing

Base case:

$$s_{\text{best}}(i, i+1) = \max_{\ell}[s(i, i+1, \ell)]$$

Pick best label

General case:

$$s_{\text{best}}(i, j) = \max_{\ell}[s(i, j, \ell)] + \max_{k}[s_{\text{best}}(i, k) + s_{\text{best}}(k, j)]$$

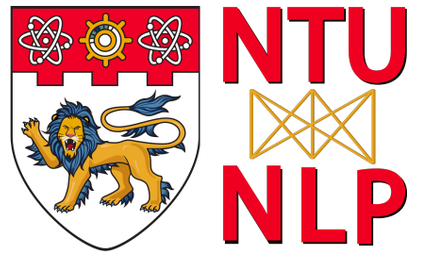
Pick best label

Pick best split point

Bottom Up

Source: Stern et al, 2017

Max-Margin Training



(Stern et al, 2017)

Want $s_{\text{tree}}(T^*) > s_{\text{tree}}(T)$ for all $T \neq T^*$

Require larger margin for higher loss:

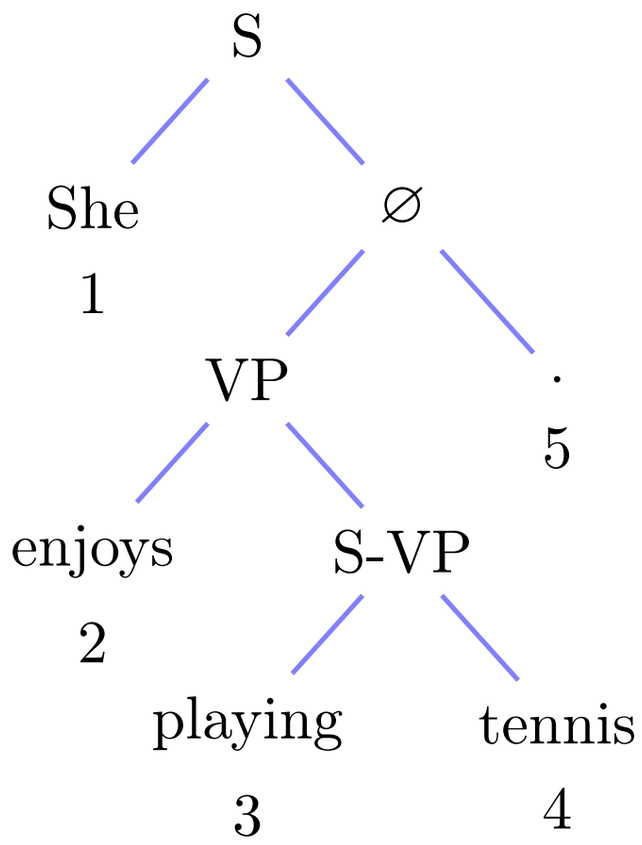
$$s_{\text{tree}}(T^*) \geq \Delta(T, T^*) + s_{\text{tree}}(T)$$

Use hinge penalty function:

$$\max\left(0, \Delta(\hat{T}, T^*) - s_{\text{tree}}(T^*) + s_{\text{tree}}(\hat{T})\right)$$

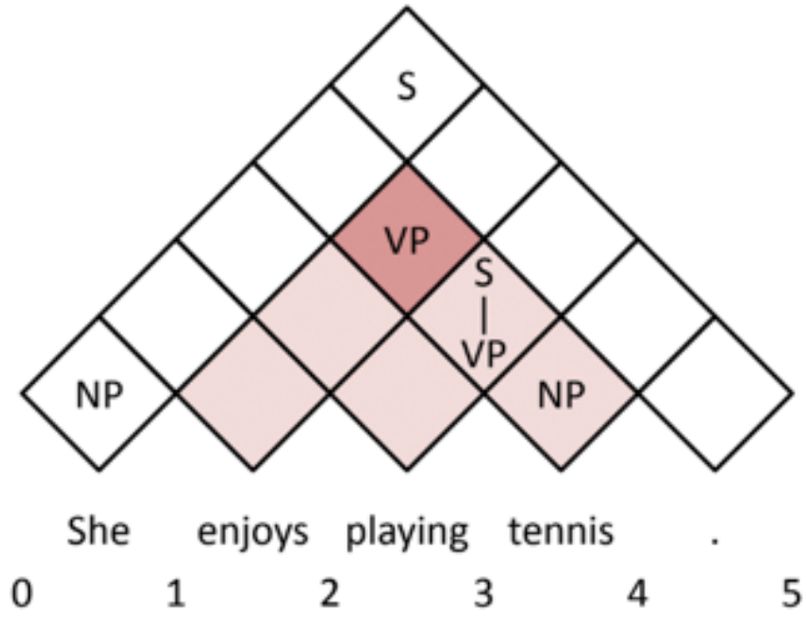
Source: Stern et al, 2017

Constituency Parsing



Bottom-up Parsing

Chart-based Neural Parser

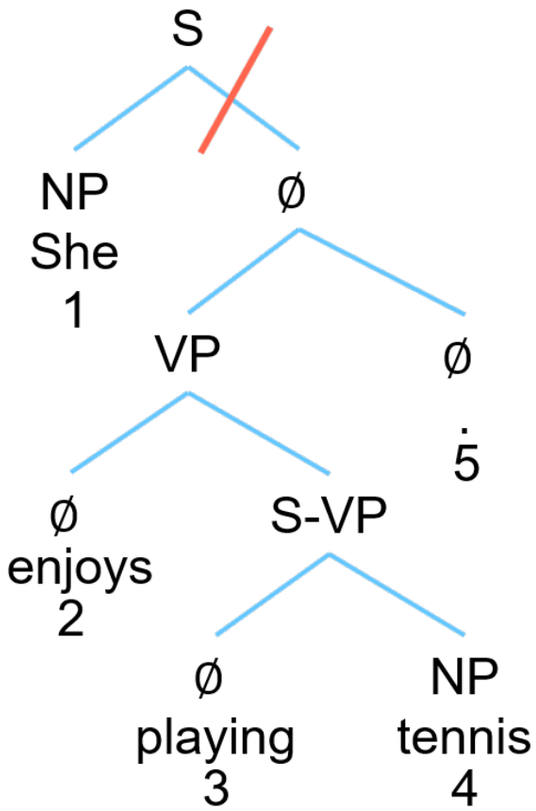


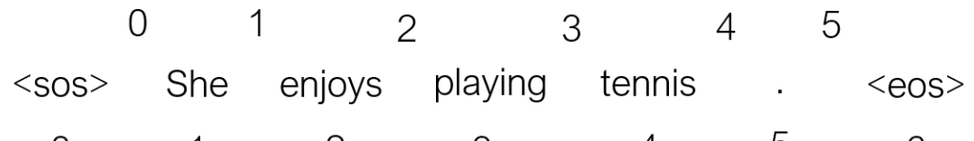
- High computational cost.
  Complexity is $\mathcal{O}(n^3)$
- Complicated loss function
  $\max(0, \max_T[s(T) + \Delta(T, T^*)] - s(T^*))$.
  Requires structured inference

Kitaev and Klein [2018]

Constituency Parsing

Top-down Greedy Parsing



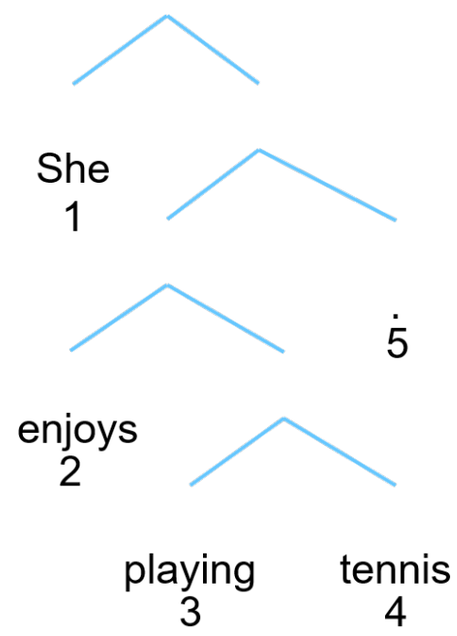


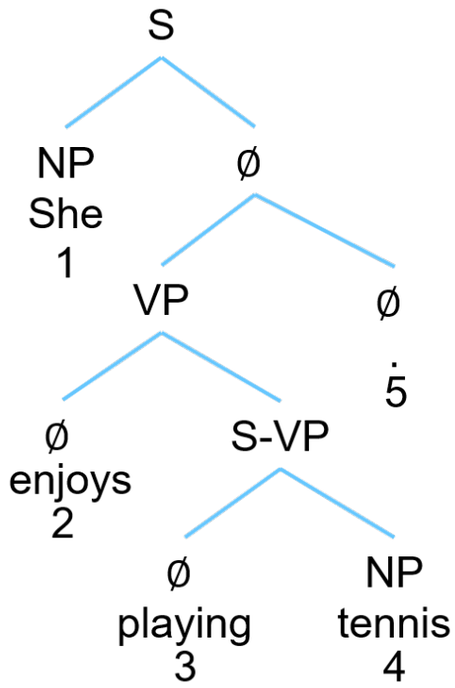
Splittings (boundary-based) $\mathcal{SP}(T) = \{(0,5)\rightarrow 1,\ (1,5)\rightarrow 4,\ (1,4)\rightarrow 2,\ (2,4)\rightarrow 3\}$

- Convert from token-based splitting $(i,j)\rightarrow(k,k+1)$ into boundary-based splitting $(i-1,j)\rightarrow k$

Constituency Parsing

Top-down Greedy Parsing

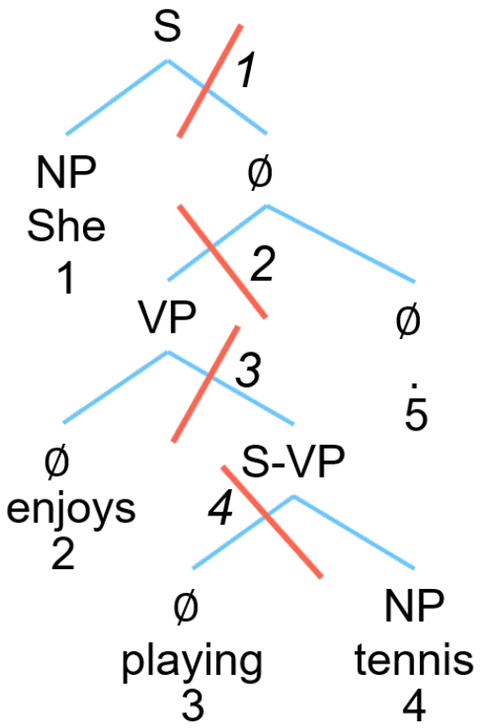




$$
\begin{aligned}
P_\theta(T|\boldsymbol{x}) &= P_\theta(L(T), \mathcal{SP}(T)|\boldsymbol{x}) \\
&= P_\theta(L(T)|\mathcal{SP}(T), \boldsymbol{x}) P_\theta(\mathcal{SP}(T)|\boldsymbol{x})
\end{aligned}
$$

Constituency Parsing

Top-down Greedy Parsing

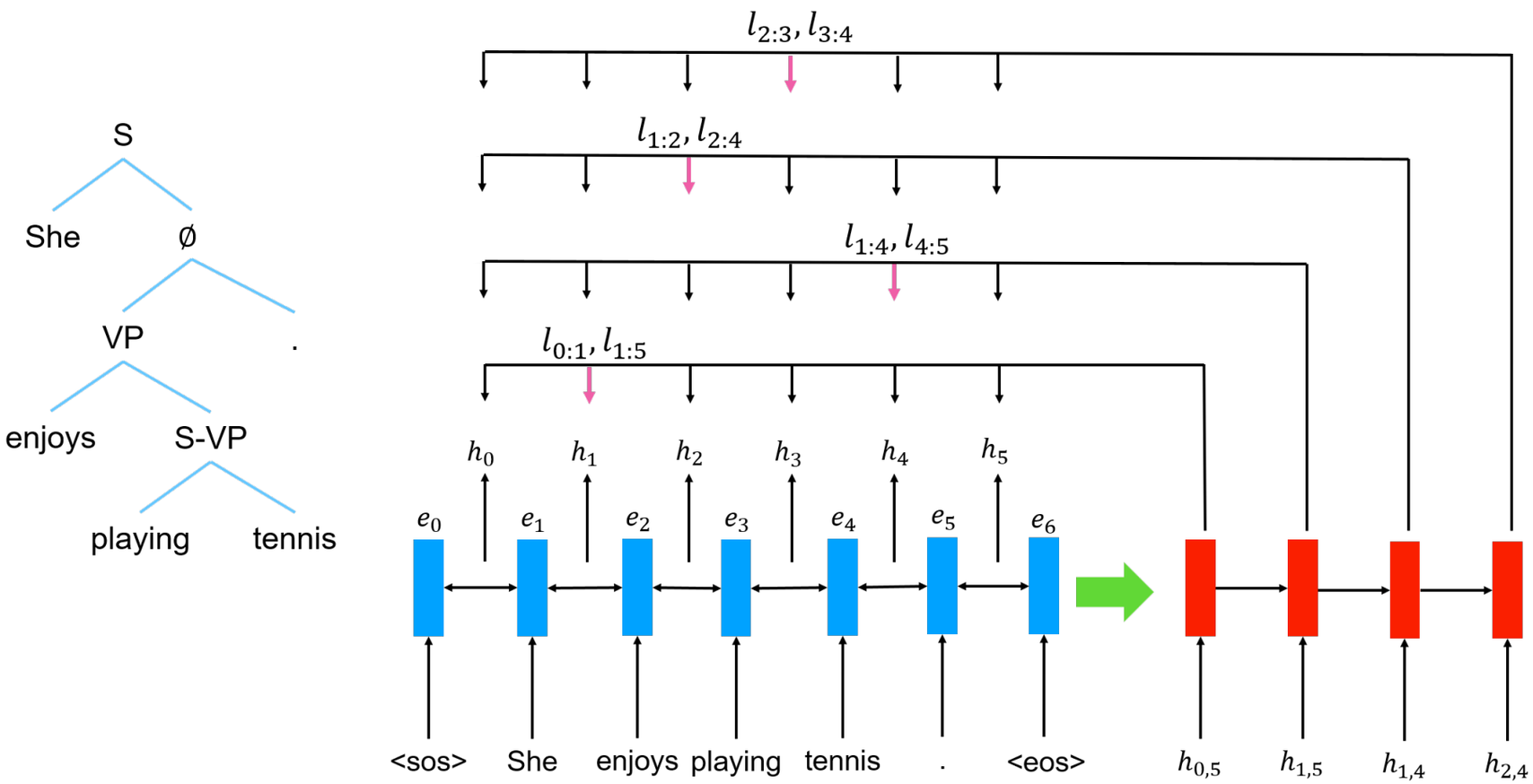


$$P_\theta(\mathcal{SP}(T)|\boldsymbol{x}) = \prod_{y_t \in \mathcal{SP}(T)} P_\theta(y_t|y_{<t}, \boldsymbol{x})$$

$$= \prod_{t=1}^{|\mathcal{SP}(T)|} P_\theta((i_t, j_t) \rightarrow k_t | ((i, j) \rightarrow k)_{<t}, \boldsymbol{x})$$

Constituency Parsing

Top-down Greedy Parsing



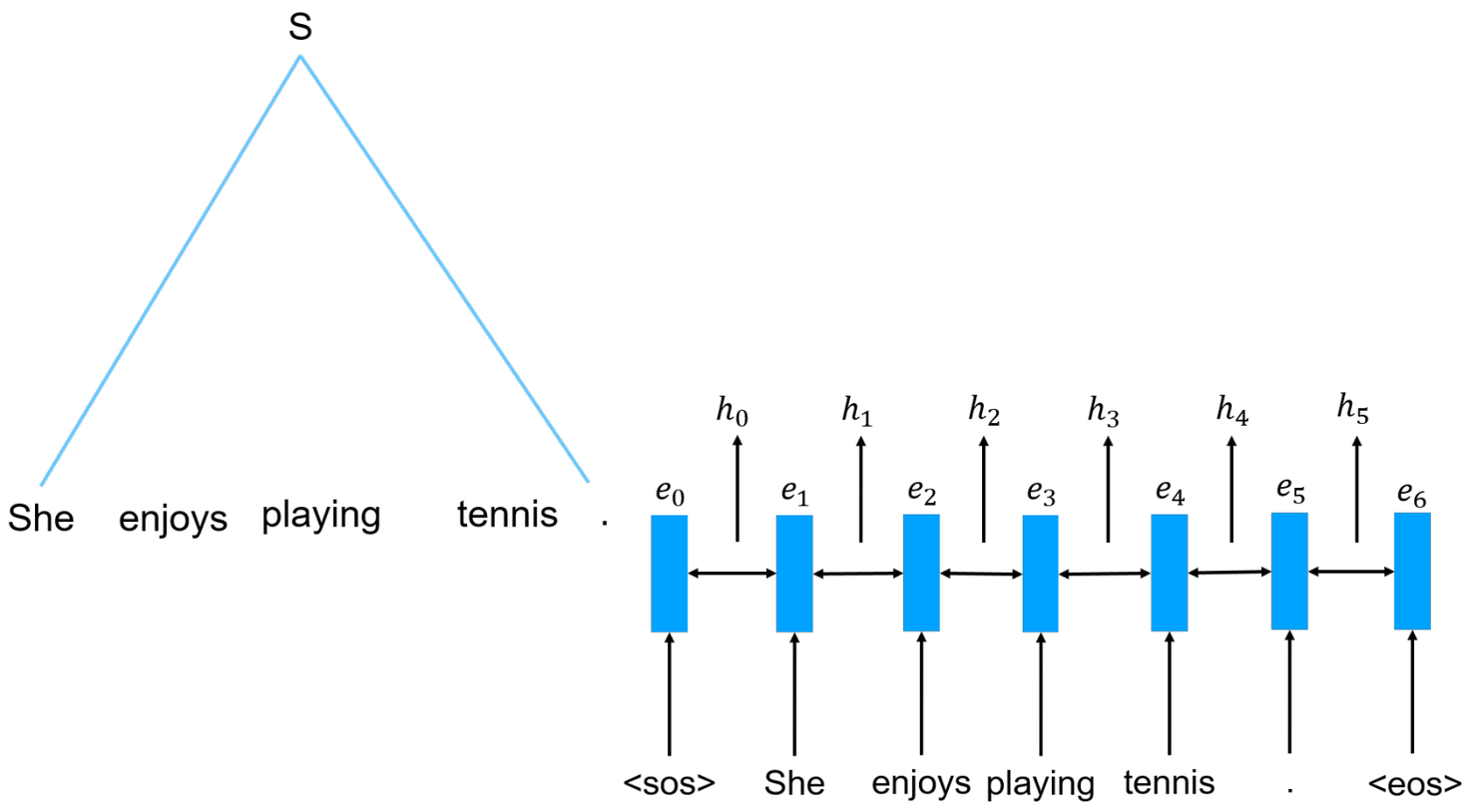
$$P_\theta(\mathcal{SP}(T)|\boldsymbol{x}) = \prod_{y_t \in \mathcal{SP}(T)} P_\theta(y_t|y_{<t}, \boldsymbol{x})$$

$$= \prod_{t=1}^{|\mathcal{SP}(T)|} P_\theta((i_t, j_t) \rightarrow k_t | ((i, j) \rightarrow k)_{<t}, \boldsymbol{x})$$

Constituency Parsing

Top-down Greedy Decoding

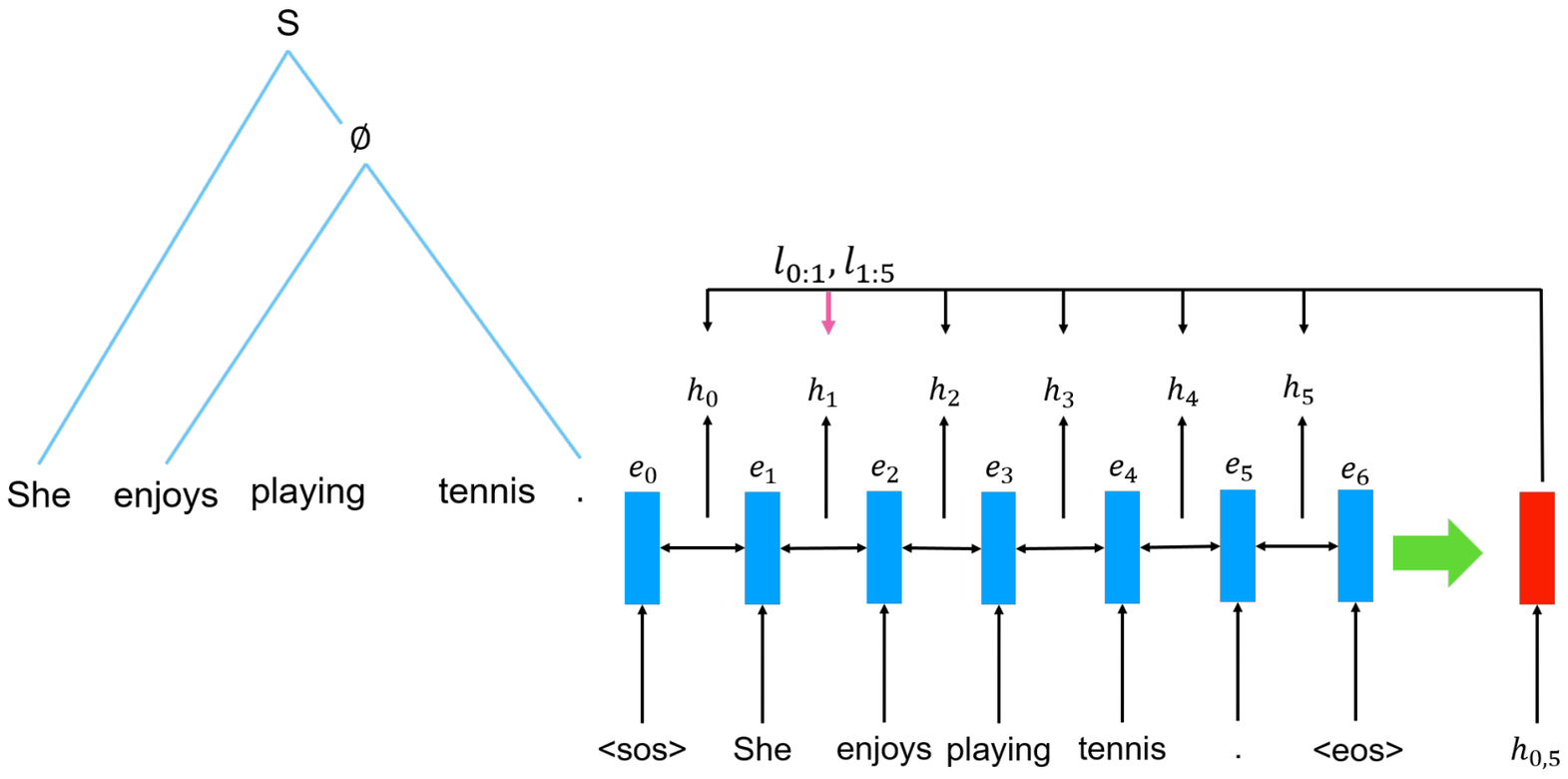
$$s = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{5} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$



Constituency Parsing

Top-down Greedy Decoding

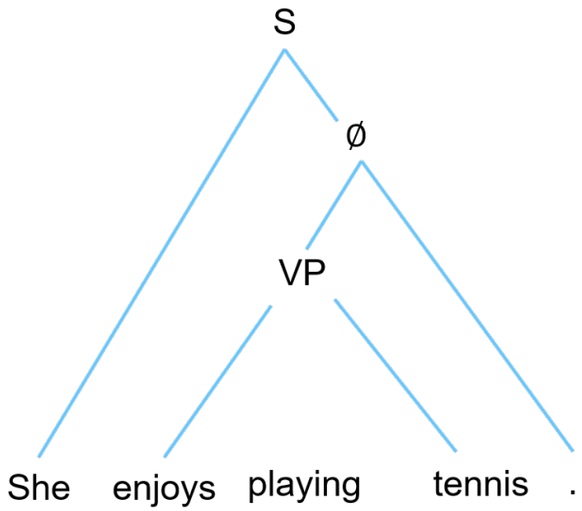
$$\boldsymbol{S} = \begin{bmatrix} \mathbf{0} \\ \mathbf{5} \end{bmatrix}, \begin{bmatrix} \mathbf{1} \\ \mathbf{5} \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

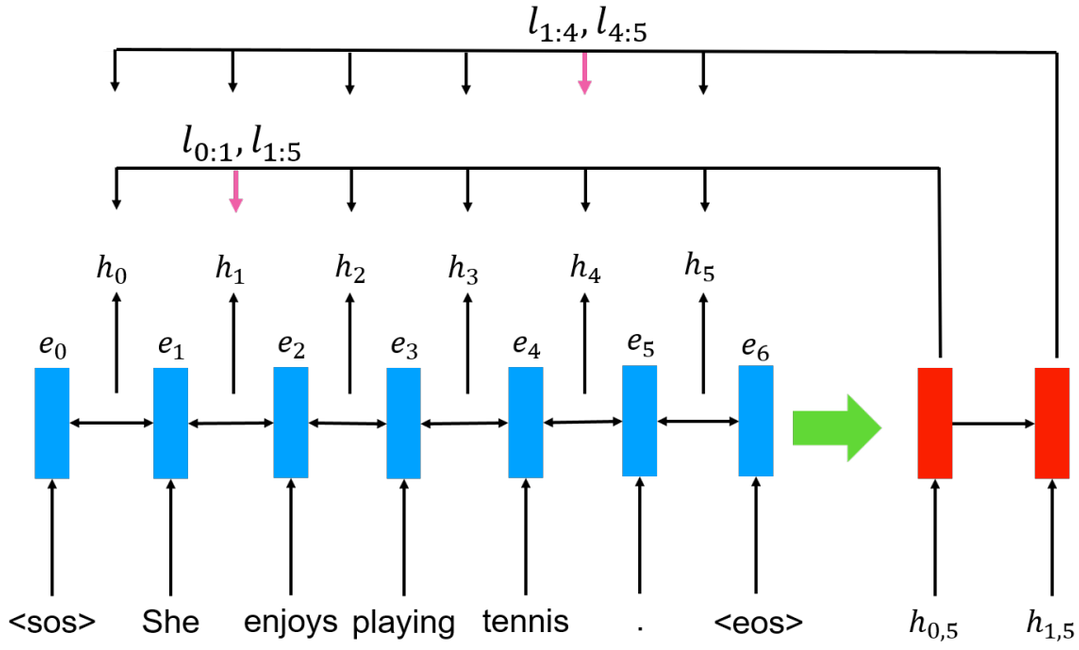


Constituency Parsing

Top-down Greedy Decoding

$$S = \begin{bmatrix} \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{5} & \mathbf{5} & \mathbf{4} & \mathbf{0} \end{bmatrix}$$

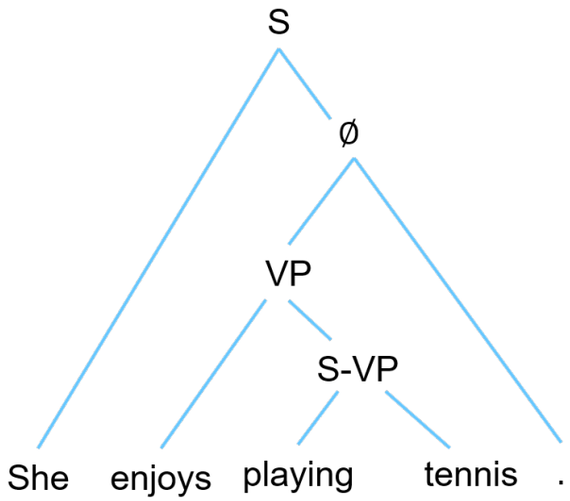


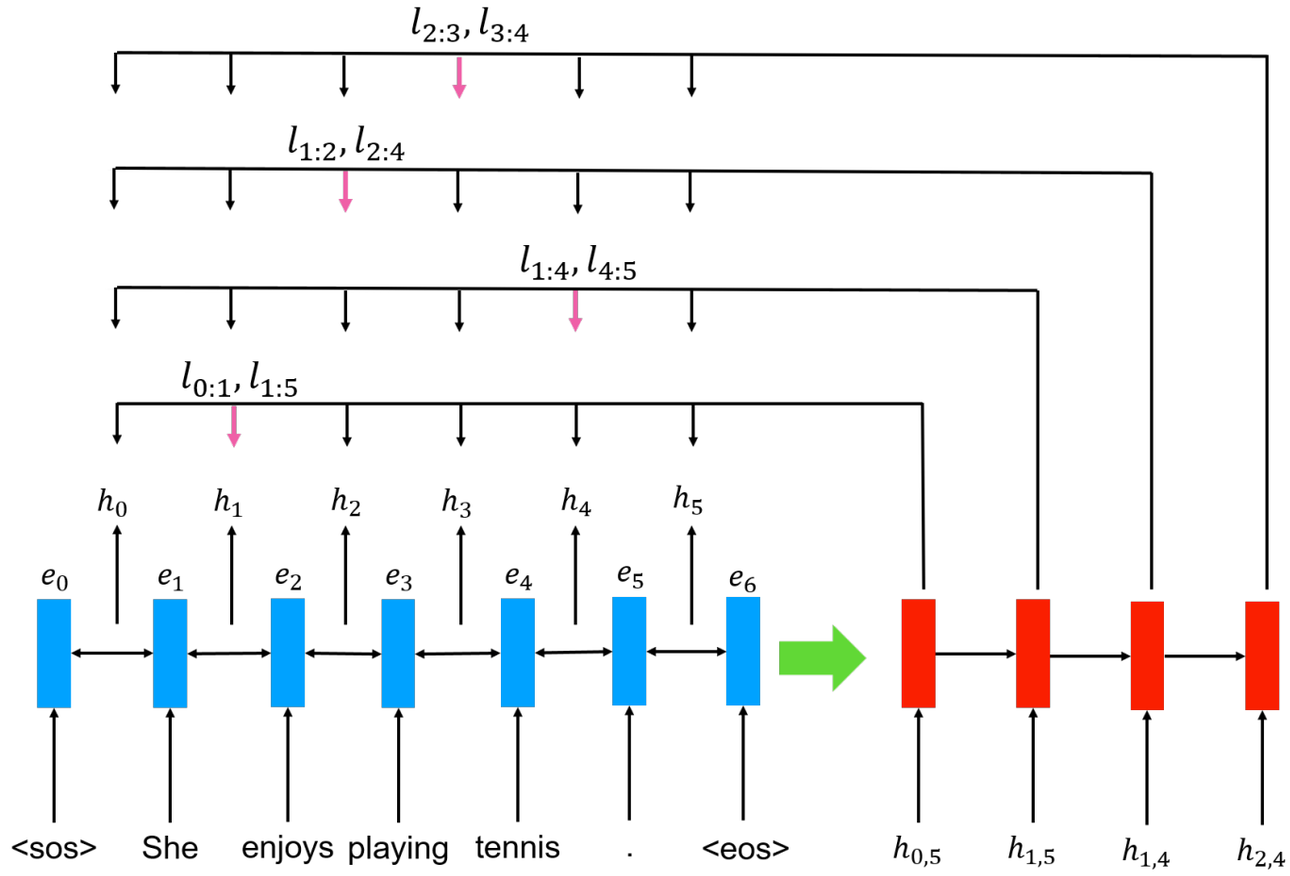


Constituency Parsing

Top-down Greedy Decoding

$$S = \begin{bmatrix} \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{2} \\ \mathbf{5} & \mathbf{5} & \mathbf{4} & \mathbf{4} \end{bmatrix}$$





Constituency Parsing

Results (English)

Model	LR	LP	F1
Top-Down Inference			
Stern et al. [2017a]	93.20	90.30	91.80
Shen et al. [2018]	92.00	91.70	91.80
Nguyen et al. [2020]	92.91	92.75	92.78
Our Model	93.90	93.63	93.77
CKY/Chart Inference			
Gaddy et al. [2018]	91.76	92.41	92.08
Kitaev and Klein [2018]	93.20	93.90	93.55
Zhang et al. [2020]	93.3	94.1	93.7
Wei et al. [2020]	93.84	93.58	93.71
Other Approaches			
Gómez and Vilares [2018]	-	-	90.7
Liu and Zhang [2017]	-	-	91.8
Stern et al. [2017b]	92.57	92.56	92.56
Zhou and Zhao [2019]	93.64	93.92	93.78

Table: Results for single models (no pre-training).

Model	F1
Nguyen et al. [2020]	95.5
Our model	95.7
Kitaev et al. [2019]	95.6
Zhang et al. [2020]	95.7
Wei et al. [2020]	95.8
Zhou and Zhao [2019]	95.8

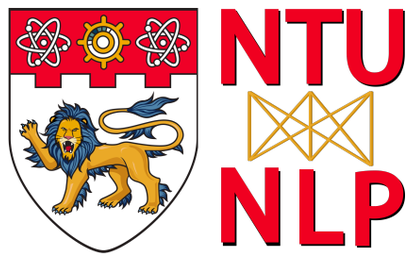
Table: Results for pretrained models.

Constituency Parsing

Model	Basque	French	German	Hungarian	Korean	Polish	Swedish
Bjorkelund et al. [2014]	88.24	82.53	81.66	91.72	83.81	90.50	85.50
Coavoux and Crabbé [2017]	88.81	82.49	85.34	92.34	86.04	93.64	84.0
Kitaev and Klein [2018]	89.71	84.06	87.69	92.69	86.59	93.69	84.45
Nguyen et al. [2020]	90.23	82.20	84.91	91.07	85.36	93.99	86.87
Our Model	89.74	84.12	85.21	92.84	86.72	92.10	85.81

Table: Results on SPMRL test sets (without any pre-training).

Dialogue Systems



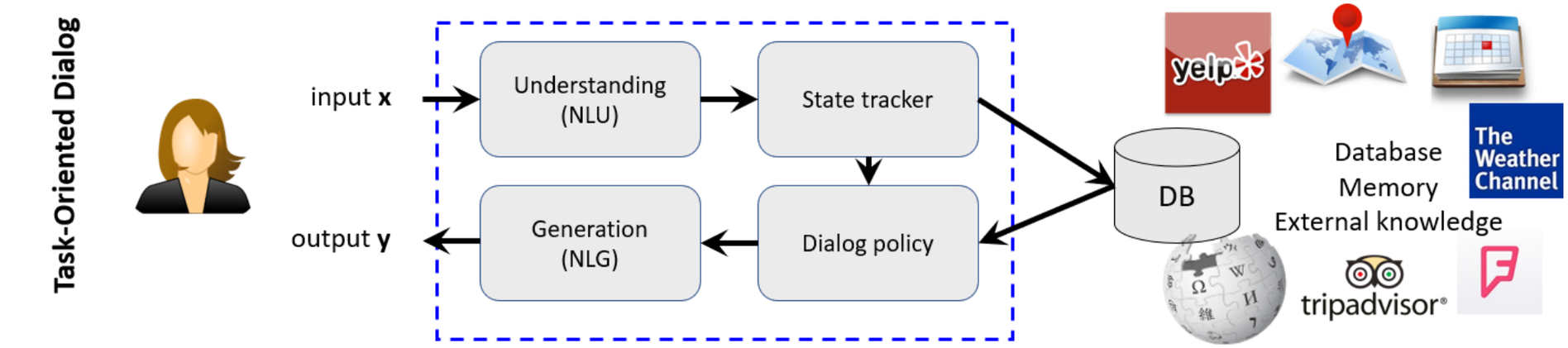
Turn	Dialogue
1	*usr:* Good morning!
	agt: Good morning, I hope you had a great weekend.
2	*usr:* not really.
	agt: You sound depressed. What can I do for you?
3	*usr:* I'd like to know where sales are lagging behind our forecast?
	agt: The worst region is [country], where sales are 15% below projections.
4	*usr:* Do you know why?
	agt: The forecast for [product] growth was overly optimistic.
5	*usr:* How can we turn this around?
	agt: Here are the 10 customers in [country] with the most growth potential, per our CRM model.
6	*usr:* Can you set up a meeting with the CTO of [company]?
	agt: Yes, I've set up a meeting with [person name] for next month when you are in [location].
7	*usr:* Thanks!

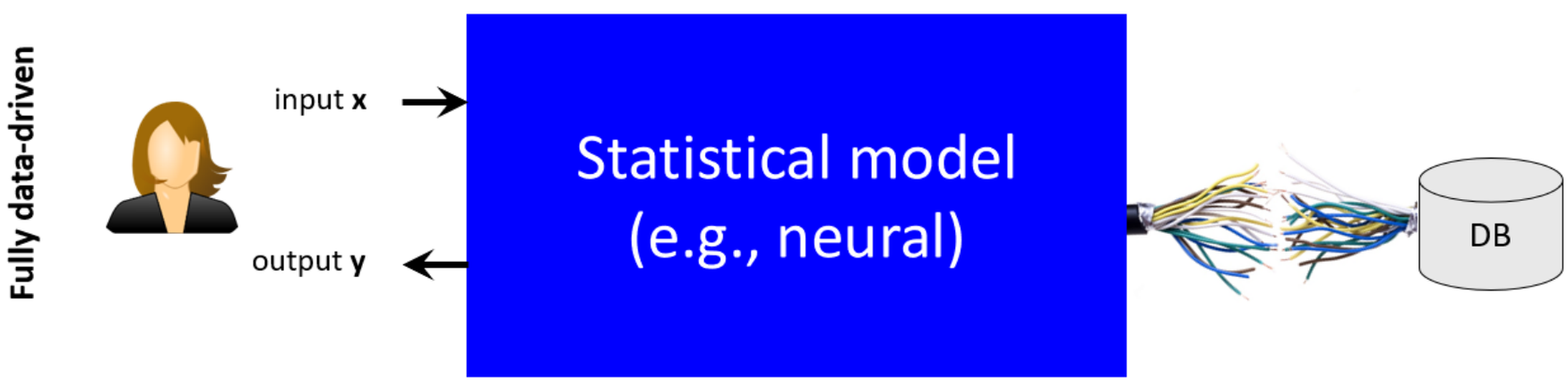
Source: Gao et al (2019)

Dialogue Systems

NTU NLP

Two Approaches:



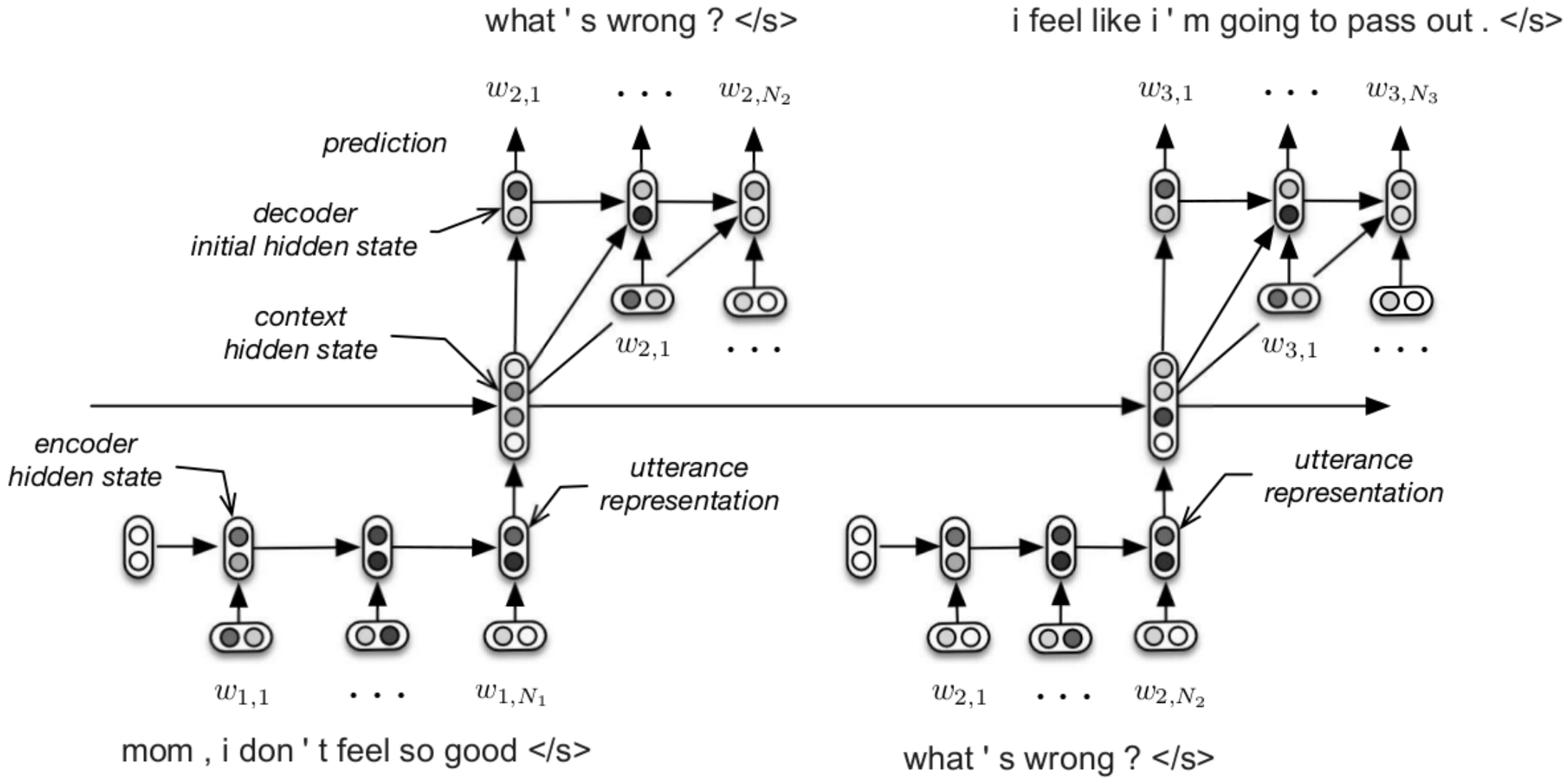


Source: Gao et al (2019)

Seq2Seq Variants

NTU NLP

Hierarchical recurrent encoder-decoder (HRED)



Source: Serban et al (2015)

Summary

NTU NLP