

# Mô hình seq2seq và ứng dụng trong sinh văn bản (seq2seq and application for text generation)

AI Academy Vietnam

- TS. Phan Việt Anh
- Nhận bằng TS năm 2018 tại Viện Khoa học Công nghệ tiên tiến Nhật Bản (JAIST)
- Lĩnh vực nghiên cứu: Trí Tuệ Nhân Tạo, Phần mềm an toàn, Xử lý ngôn ngữ tự nhiên, Xử lý âm thanh.
- Giảng dạy: Phần mềm an toàn, xử lý ngôn ngữ tự nhiên, học máy.
- Dự án: STT, TTS, Tổng hợp và phân tích thông tin trên mạng xã hội;

# Nội dung trình bày

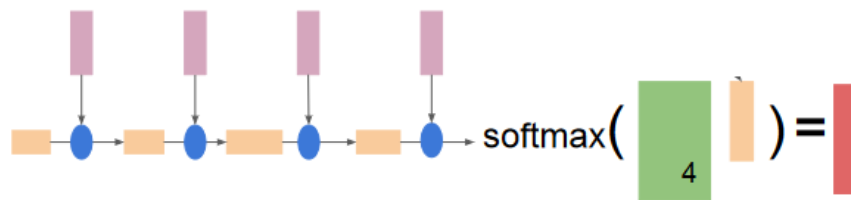
---

- Mô hình seq2seq
- Giới thiệu về sinh văn bản
  - Sinh văn bản
  - Dịch máy
- Kỹ thuật Attention
- Dịch máy với seq2seq

# Mô hình Seq2Seq

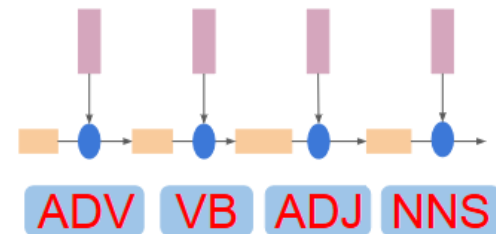
- Áp dụng vào các bài toán: sinh token đầu ra tại mỗi bước
  - Xác định từ loại (POS)
  - Trích xuất thực thể (NER)
  - Dịch máy (machine translation)
  - Nhận dạng âm thanh (speech recognition)
  - Xác định tiêu đề cho bức ảnh (image captioning)

Only use neural nets



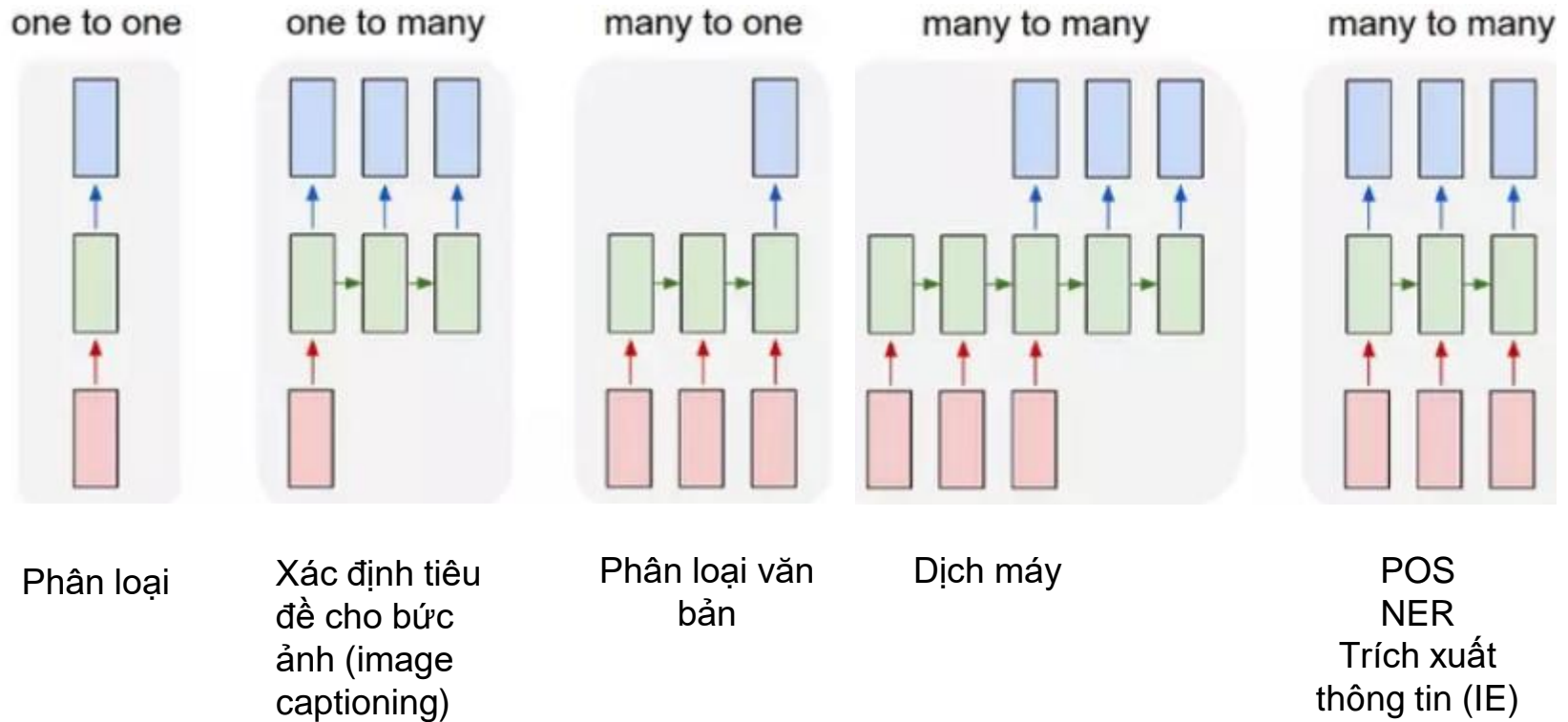
RNN cho mô hình ngôn ngữ, phân loại câu

Only use neural nets



RNN Cho gán nhãn từ loại

# Một số mô hình Seq2Seq



Một số loại mô hình Seq2Seq và ứng dụng

# Dịch máy

---

## Dịch máy (Machine Translation - MT)

- Đầu vào: một câu  $x$  thuộc một ngôn ngữ (ngôn ngữ nguồn)
- Đầu ra: câu  $y$  trong ngôn ngữ khác (ngôn ngữ đích).

$x:$       *L'homme est né libre, et partout il est dans les fers*



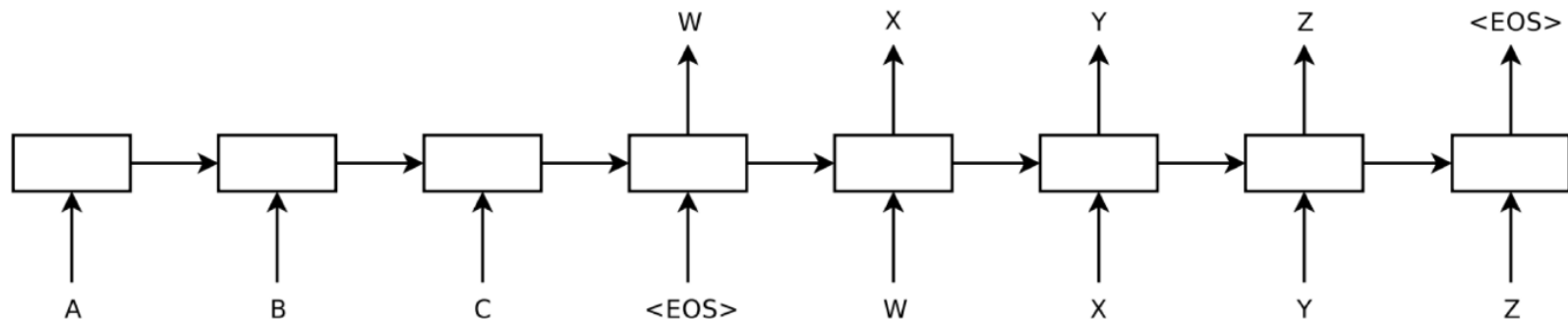
$y:$       *Man is born free, but everywhere he is in chains*

# Dịch máy sử dụng mạng nơ ron là gì?

---

- Dịch máy mạng nơ ron – (Neural Machine Translation - **NMT**): là Phương pháp dịch máy sử dụng một mạng nơ ron
- Kiến trúc mạng nơ ron được gọi là sequence-to-sequence (hay **seq2seq**) và thường là 2 RNNs.

# Mô hình mã hóa, giải mã (Encoder-decoder )

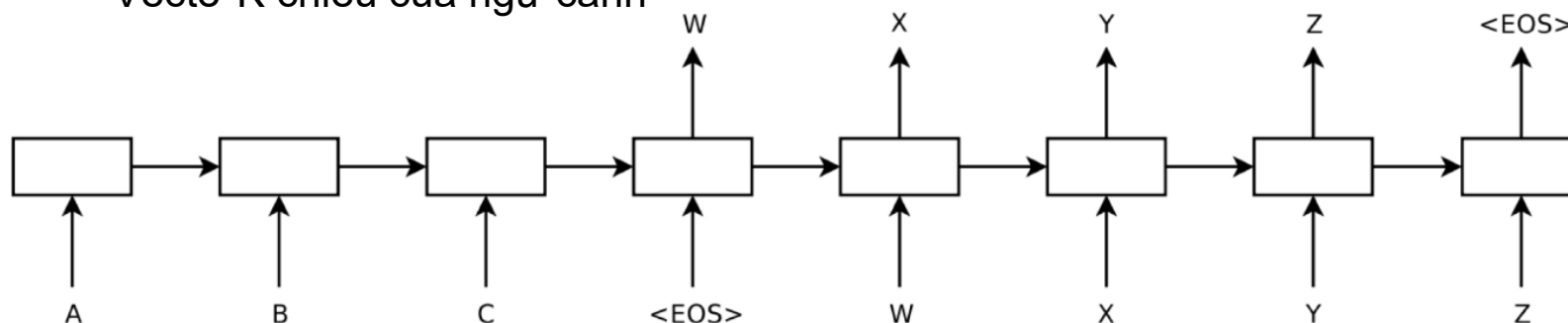


“Sequence to Sequence Learning with Neural Networks”, 2014



# Mô hình mã hóa, giải mã (Encoder-decoder )

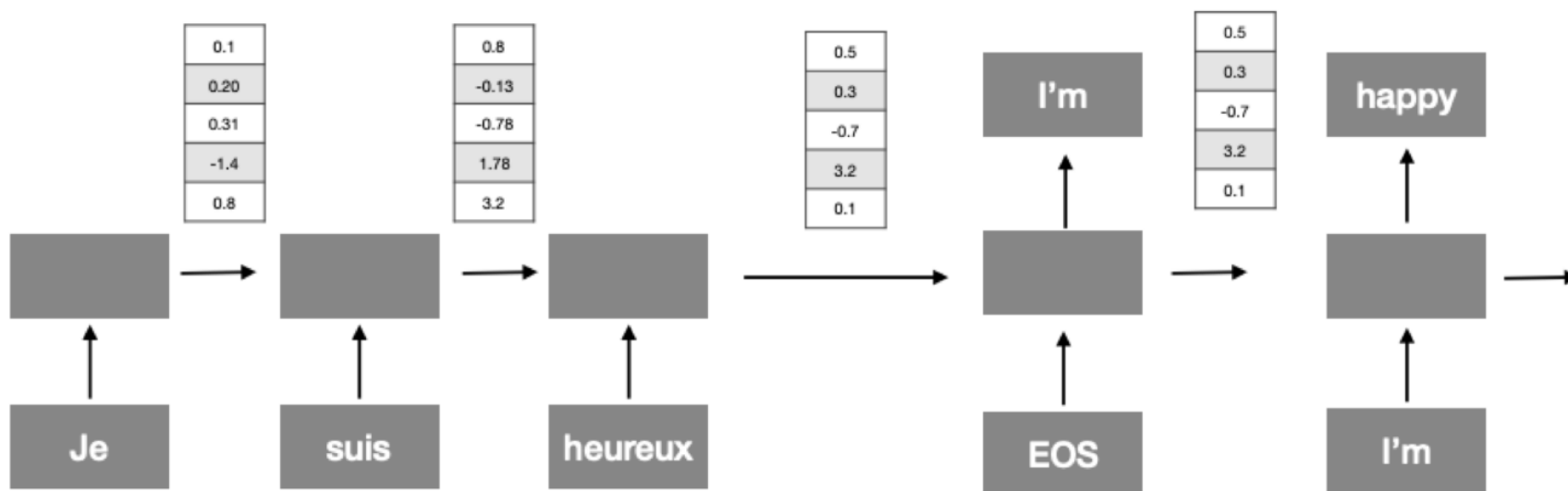
Vectơ K chiều của ngữ cảnh



Điều kiện của từ được sinh ra trong bản dịch

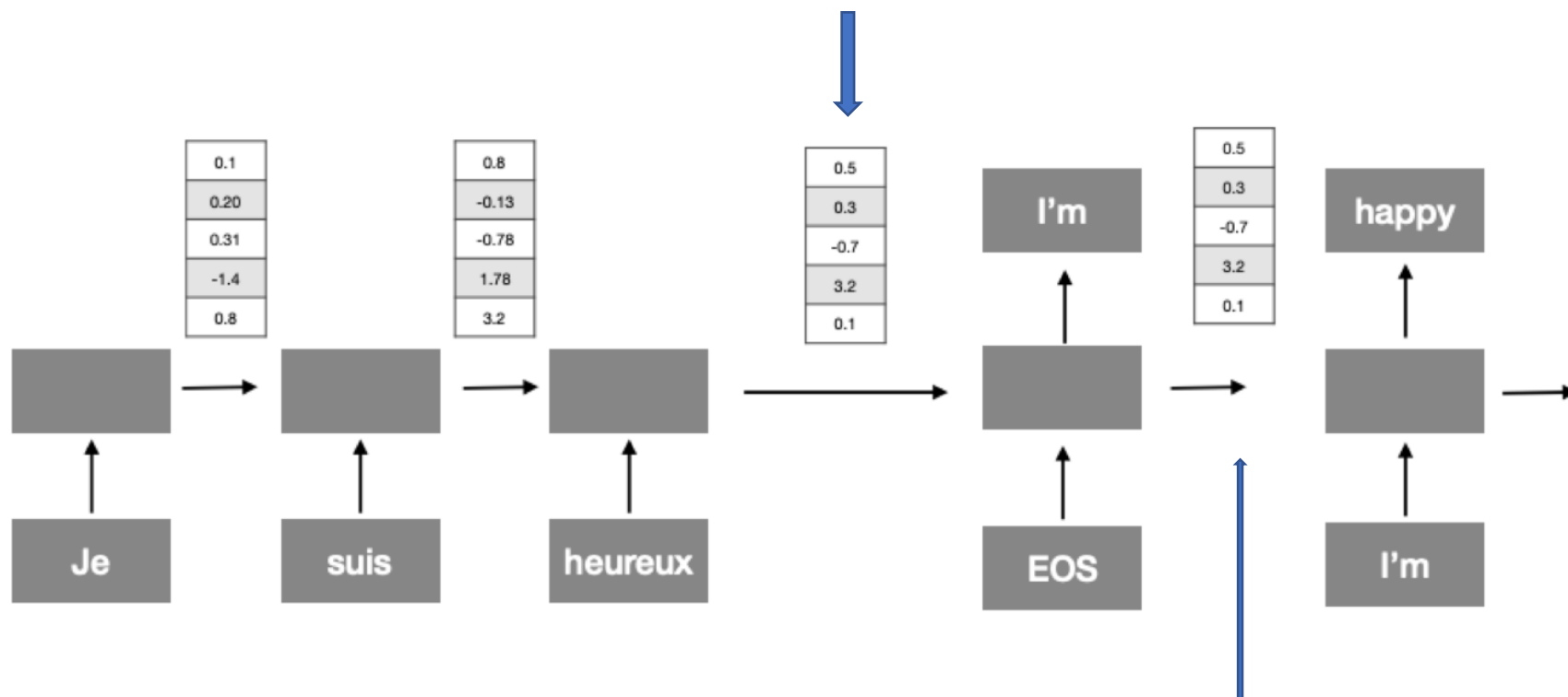
“Sequence to Sequence Learning with Neural Networks”, 2014

# Mô hình mã hóa, giải mã (Encoder-decoder)



# Mô hình mã hóa, giải mã (Encoder-decoder)

Toàn bộ đầu vào được tổng hợp trong một vector duy nhất này

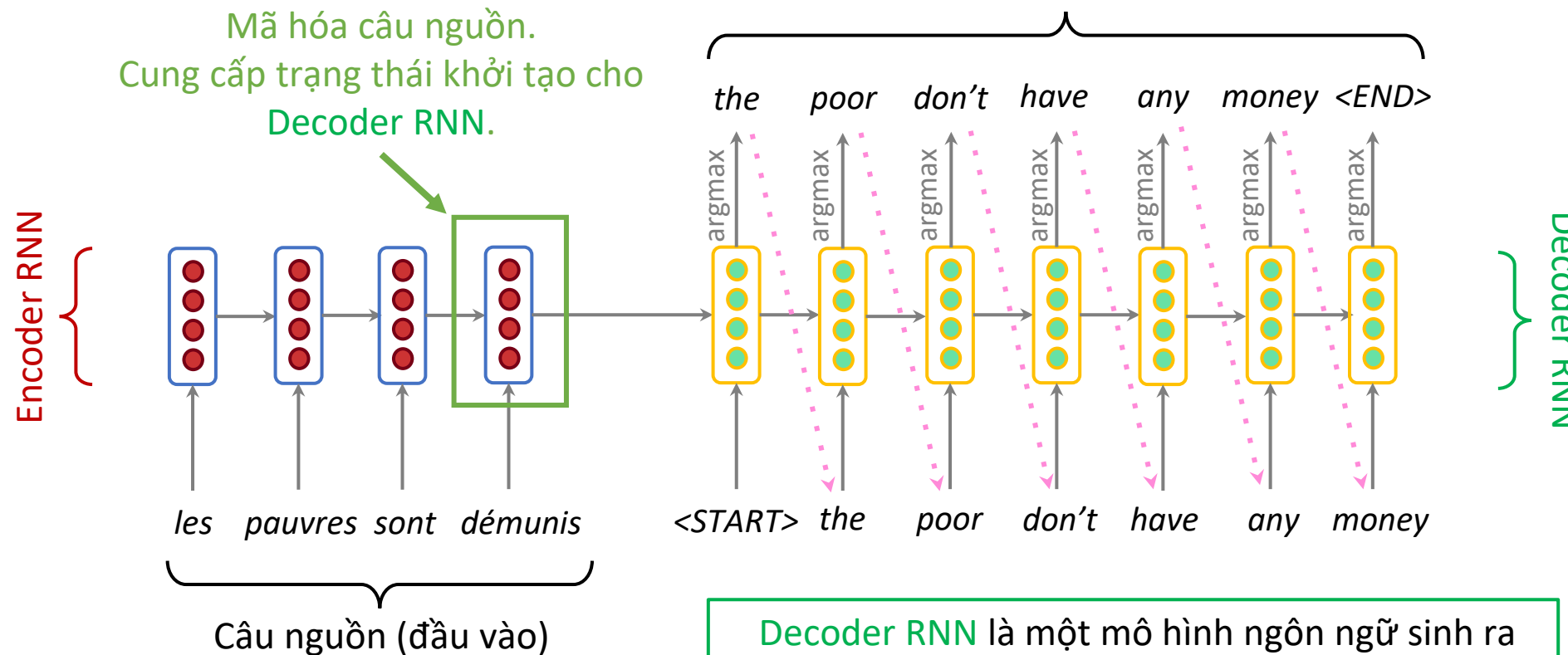


Trong mô hình seq2seq, trạng thái của bộ giải mã chỉ phụ thuộc vào trạng thái trước đó và đầu ra trước đó

# Dịch máy dựa vào mạng nơ ron (NMT)

Mô hình Seq2Seq

Câu đích (đầu ra)



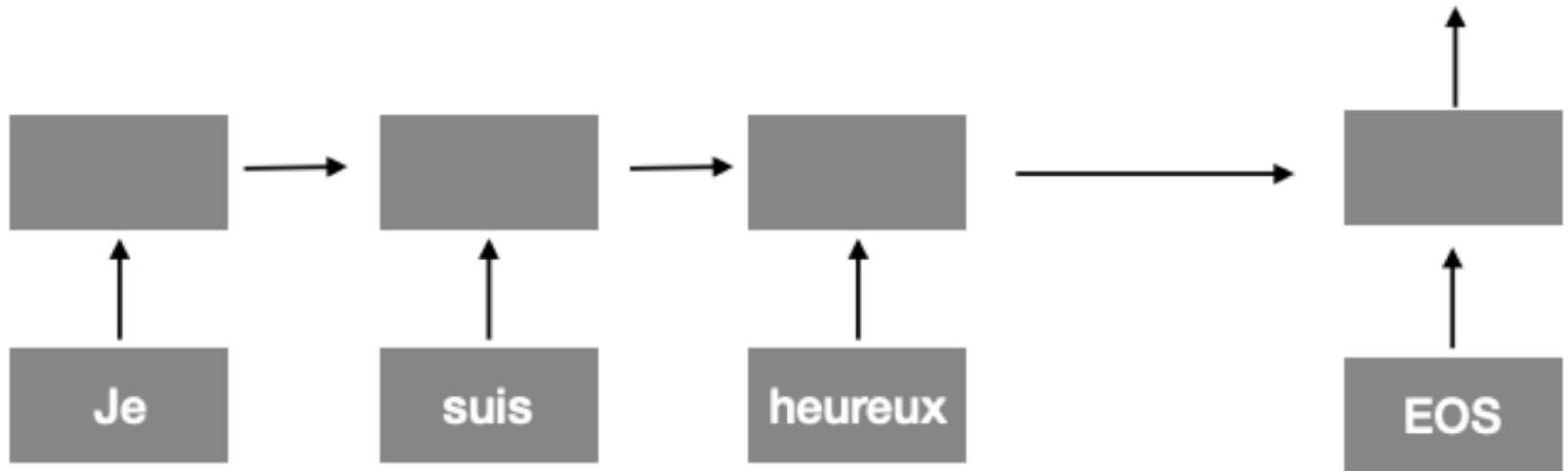
Encoder RNN sinh ra  
vector mã hóa cho câu  
nguồn.

Decoder RNN là một mô hình ngôn ngữ sinh ra  
câu đích có điều kiện trên **encoding**.

Lưu ý: Sơ đồ này thể hiện hành vi của mô hình tại  
thời điểm **test** : đầu ra của decoder được cung  
cấp như đầu vào của bước tiếp theo

# Huấn luyện

- Như trong mô hình RNN khác, chúng ta có thể huấn luyện bằng cách tối thiểu hàm loss giữa những gì chúng ta dự đoán ở mỗi bước và giá trị đúng của nó.



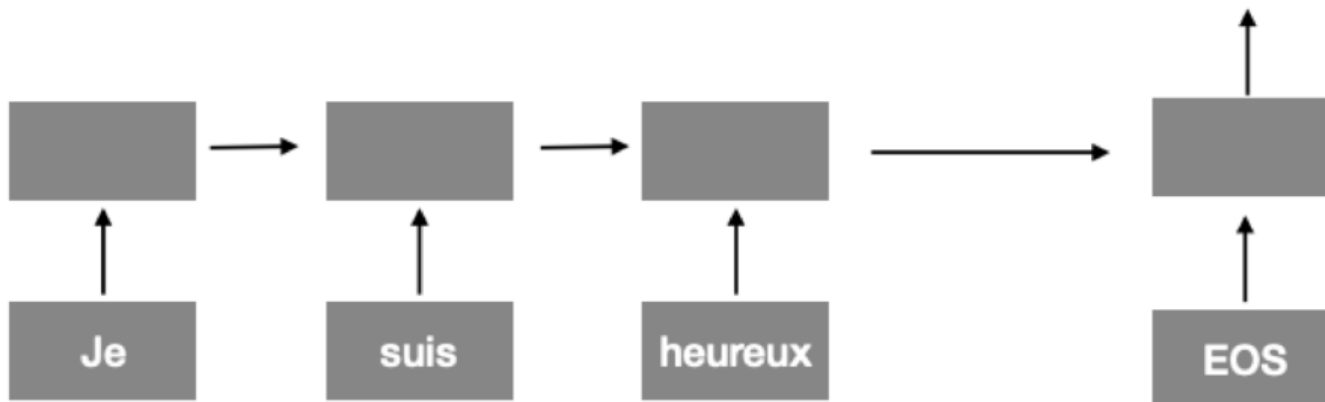
# Huấn luyện

Truth

I'm	you	are	the	...
1	0	0	0	0

Predicted

I'm	you	are	the	...
0.03	0.05	0.02	0.01	0.009



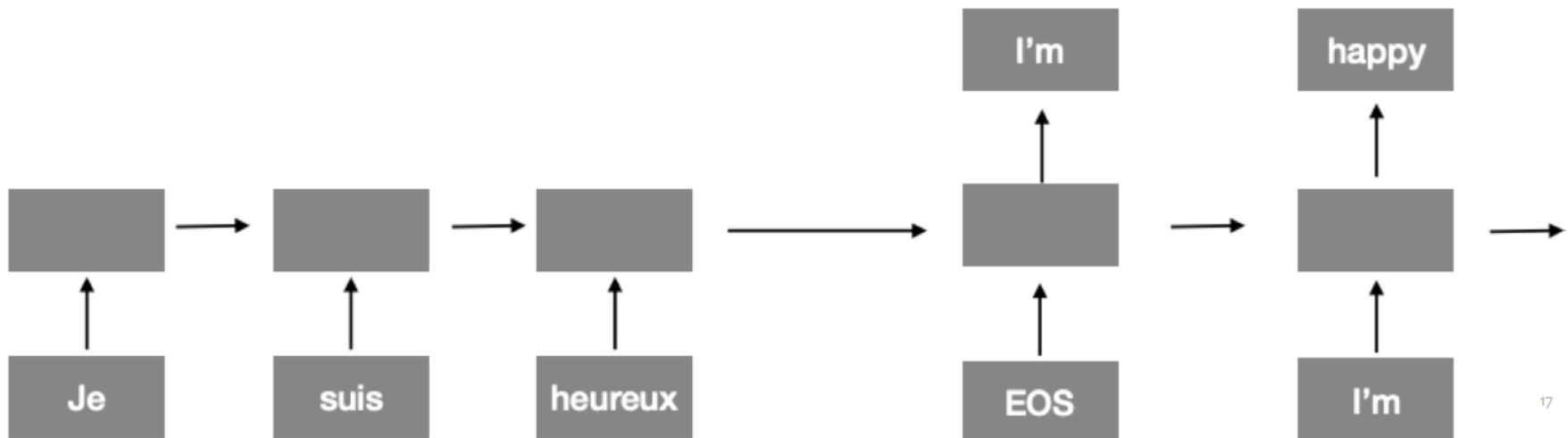
# Huấn luyện

Truth

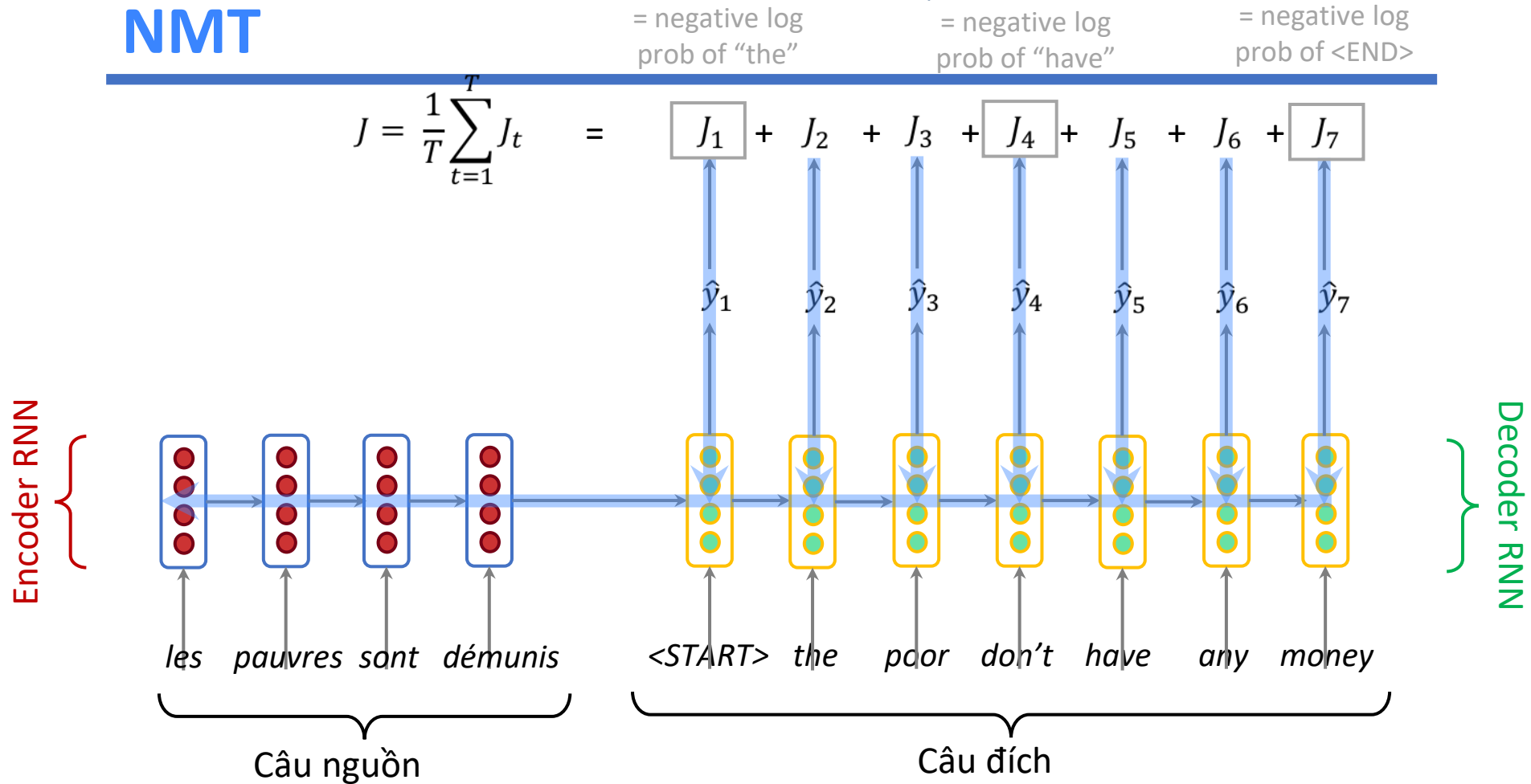
happy	great	bad	ok	...
1	0	0	0	0

Predicted

happy	great	bad	ok	...
0.13	0.08	0.01	0.03	0.009



# Huấn luyện hệ thống NMT

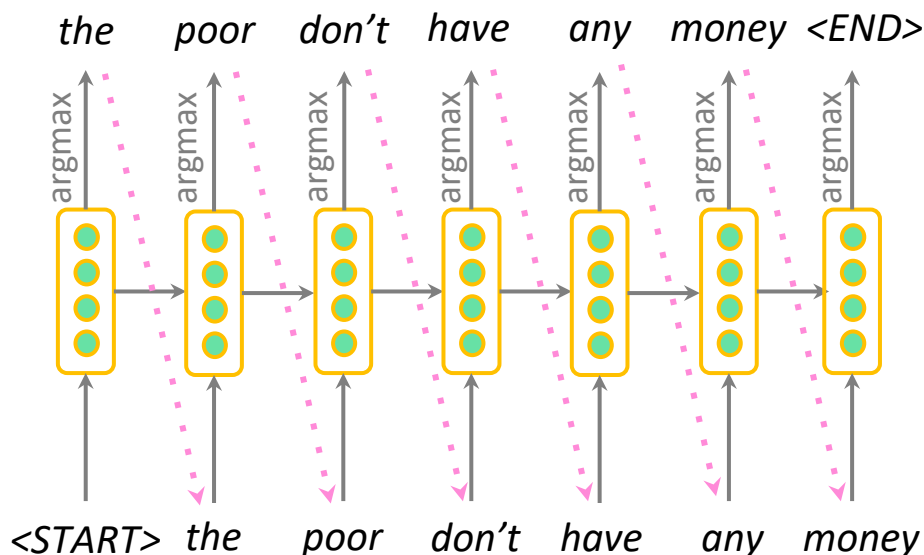


Seq2seq được tối ưu như là 1 hệ thống.  
Thuật toán lan truyền ngược được tự hiện đầu – cuối



# Giải mã tham lam (greedy decoding)?

- Sinh ra câu đích bằng cách lấy từ có xác suất lớn nhất tại mỗi bước giải mã



- Đây gọi là giải mã tham lam
- **Vấn đề đối với giải mã tham lam là gì?**

# Giải mã tham lam?

---

- Giải mã tham lam không thể quay về quyết định trước!
  - *les pauvres sont démunis (the poor don't have any money)*
  - → the \_\_\_\_\_
  - → the poor \_\_\_\_\_
  - → the poor *are* \_\_\_\_\_
- Lựa chọn tốt hơn: sử dụng beam search để tìm kiếm một vài ứng viên và lựa chọn giải pháp tốt nhất

# Giải mã dựa vào Beam search

---

$$P(y|x) = P(y_1|x) P(y_2|y_1, x) P(y_3|y_1, y_2, x) \dots, P(y_T|y_1, \dots, y_{T-1}, x)$$

- Mục tiêu là chúng ta tìm  $y$  tối ưu
- Nếu chúng ta liệt kê toàn bộ  $y \rightarrow$  lượng tính toán và lưu trữ quá lớn
  - Độ phức tạp  $O(V^T)$  trong đó  $V$  là kích thước từ vựng,  $T$  là chiều dài câu đích
- **Beam search**
  - Tại mỗi bước giải mã, giữ lại  $k$  phần dịch tốt nhất.
  - $K$  là kích thước của beam (trong thực tế thường 5, 10)
  - Dùng beam search không đảm bảo tìm được lời giải tối ưu
  - Nhưng nó hiệu quả hơn rất nhiều so với tìm kiếm tham lam

# Giải mã dựa vào Beam search: ví dụ

---

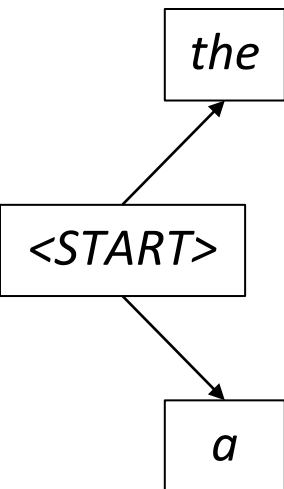
Beam size = 2

<START>

# Giải mã dựa vào Beam search: ví dụ

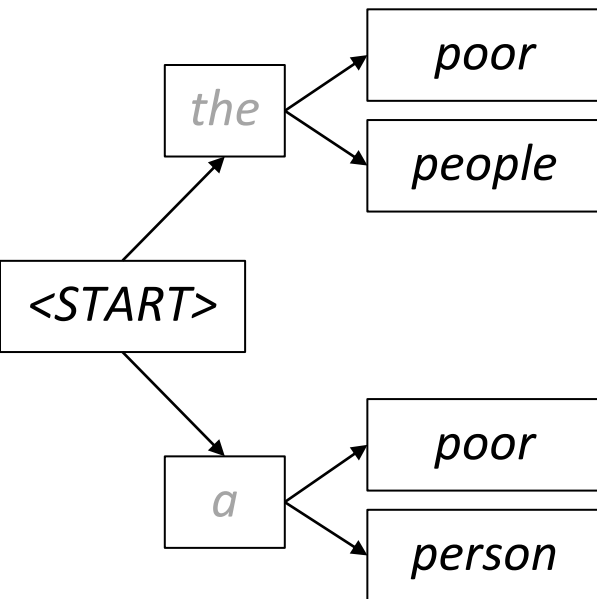
---

Beam size = 2



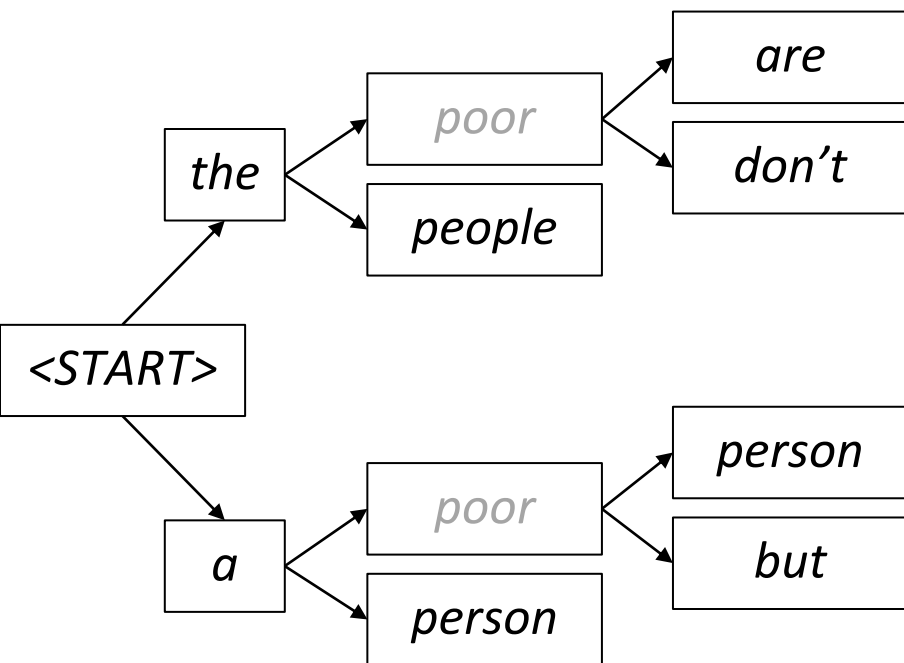
# Giải mã dựa vào Beam search: ví dụ

Beam size = 2



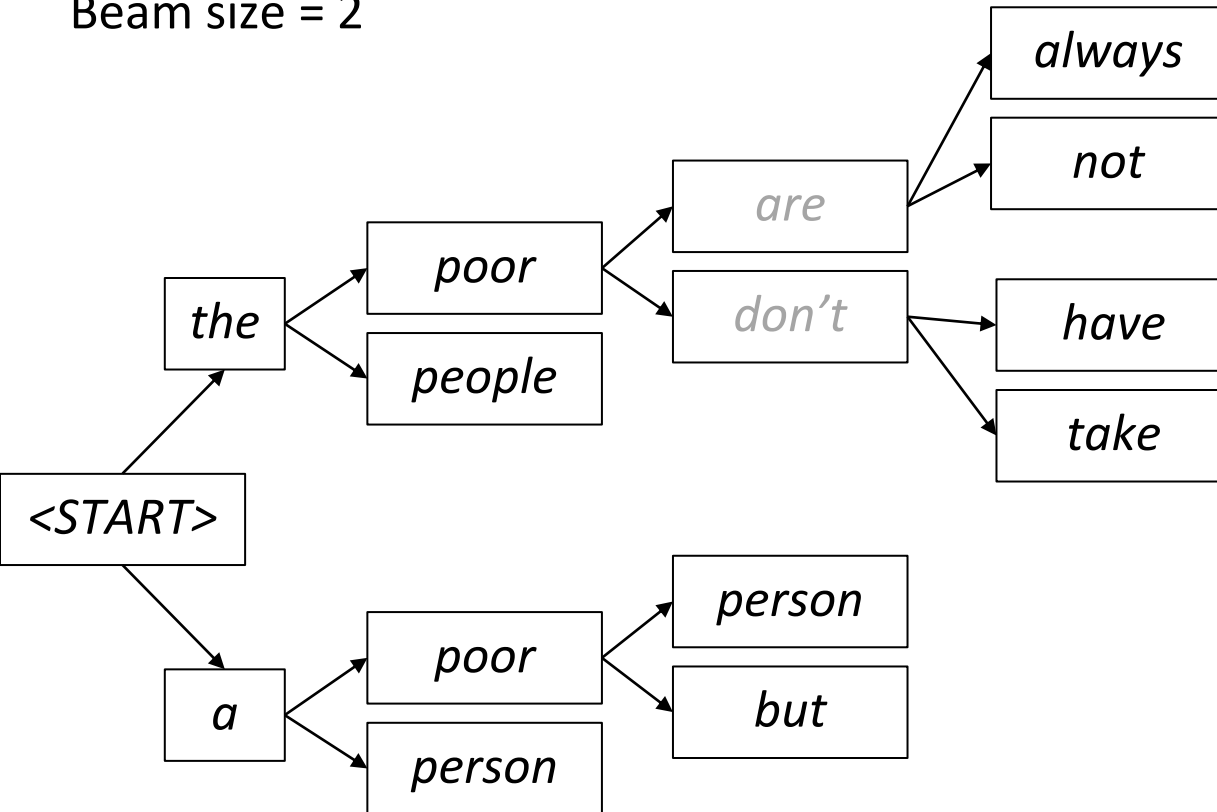
# Giải mã dựa vào Beam search: ví dụ

Beam size = 2



# Giải mã dựa vào Beam search: ví dụ

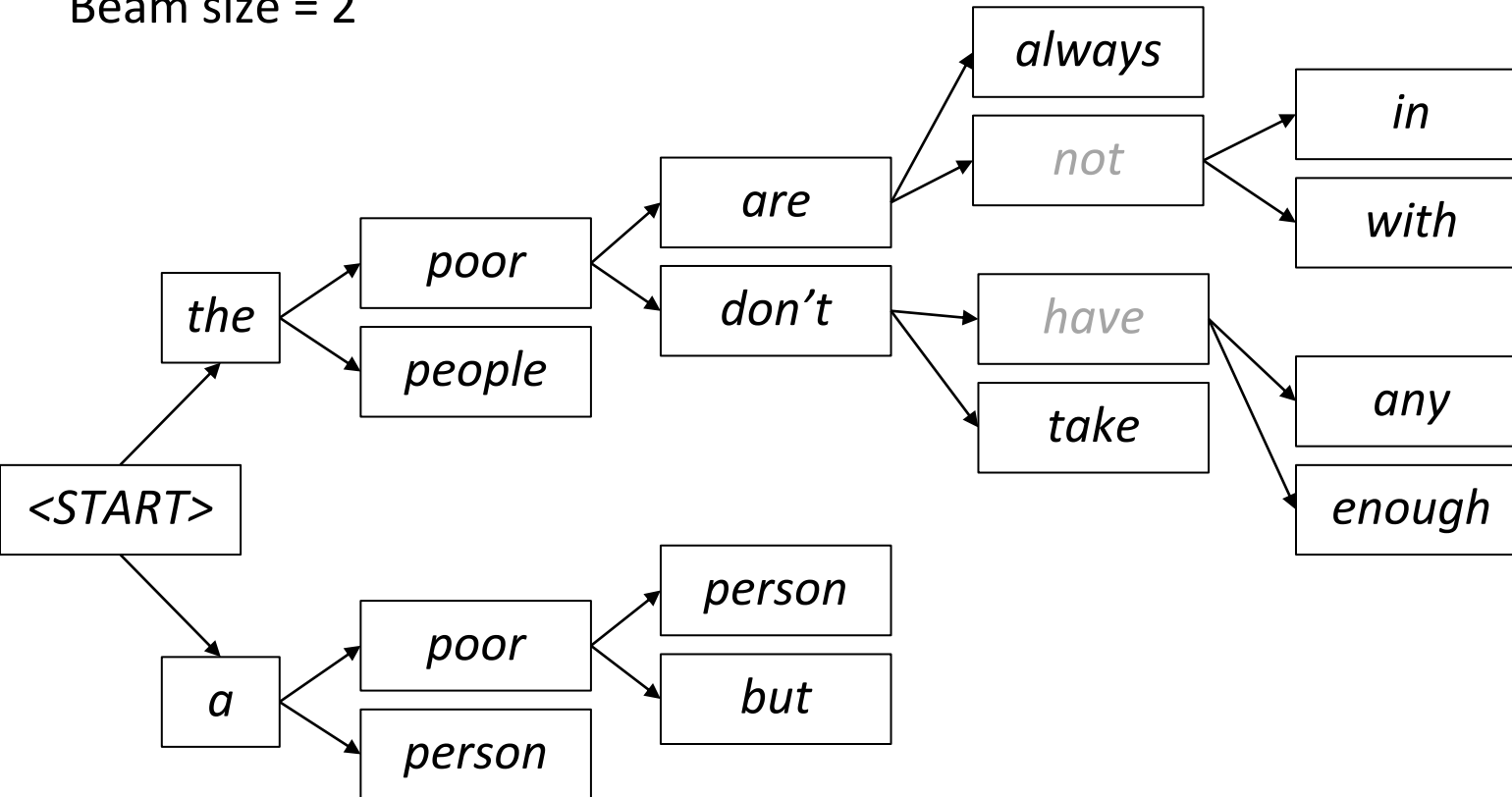
Beam size = 2





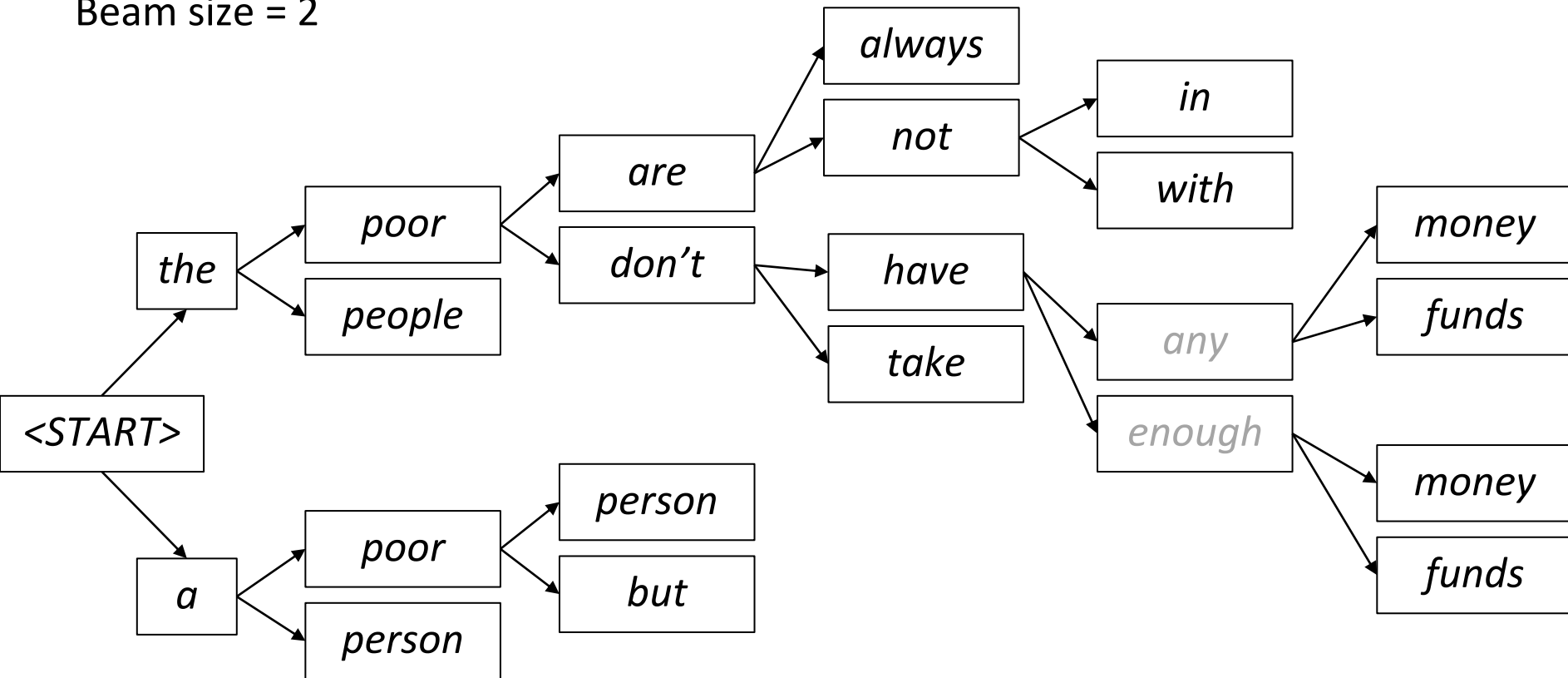
# Giải mã dựa vào Beam search: ví dụ

Beam size = 2



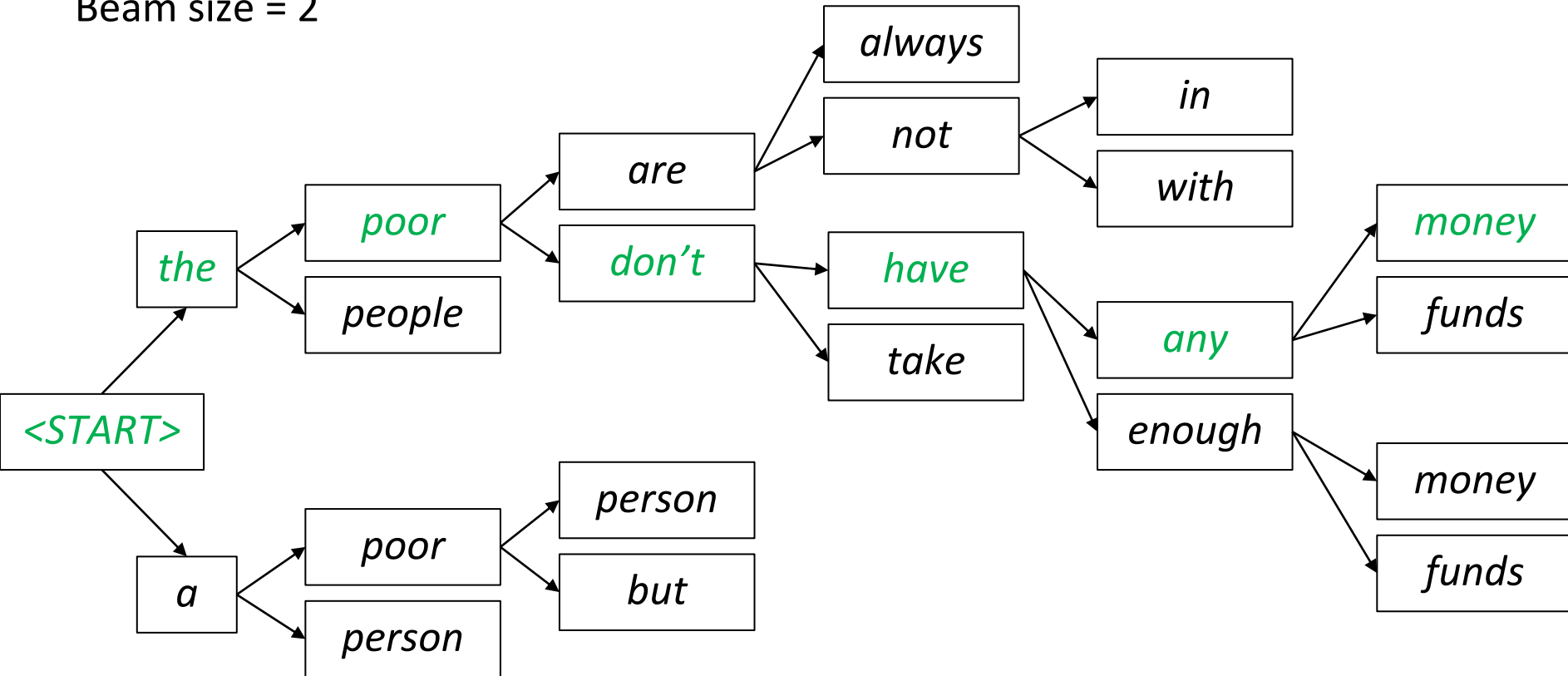
# Giải mã dựa vào Beam search: ví dụ

Beam size = 2



# Giải mã dựa vào Beam search: ví dụ

Beam size = 2



# Beam search: Tiêu chí dừng

---

- Trong giải mã tham lam, chúng ta thường giải mã cho đến khi mô hình sinh ra token `<END>`

Ví dụ : `<START> he hit me with a pie <END>`

- Trong giải mã beam search, các giả thuyết khác nhau có thể tạo ra token `<END>` ở các bước thời gian khác nhau.
  - Khi một giả thuyết sinh ra `<END>`, thì giả thuyết đó đã hoàn thành.
  - Đặt nó sang một bên và tiếp tục khám phá các giả thuyết khác thông qua beam Search.
- Thông thường, chúng tôi tiếp tục beam search cho đến khi:
  - Chúng ta đạt đến bước thời gian T (T là một số ngưỡng đã được xác định trước) hoặc
  - Chúng ta có ít nhất n giả thuyết đã hoàn thành (trong đó n là ngưỡng được xác định trước)

# Beam search: Kết thúc

- Chúng ta có danh sách ứng viên cho câu đích
- Làm thế nào để chọn câu tốt nhất
- Mỗi ứng viên  $y_1, y_2, \dots, y_t$ , điểm của nó là

$$\text{score}(y_1, \dots, y_t) = \log P_{\text{LM}}(y_1, \dots, y_t | x) = \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$$

- Vấn đề: các câu dài sẽ có điểm thấp
- Giải quyết: chuẩn hóa bởi chiều dài. Sử dụng điểm này để chọn câu tốt nhất

$$\frac{1}{t} \sum_{i=1}^t \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$$

# Ưu điểm của NMT

---

So với dịch máy thống kê (SMT), NMT có nhiều ưu điểm:

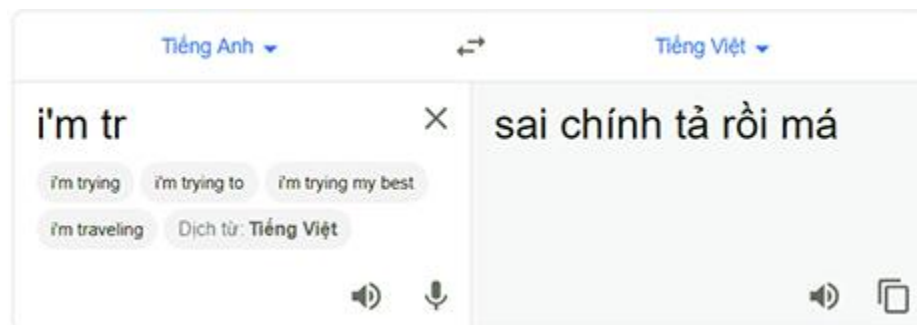
## Chất lượng cao hơn

- Trôi chảy hơn
- Sử dụng ngữ cảnh tốt hơn
- Sử dụng các cụm từ tương tự tốt hơn
- Một mạng nơ ron được tối ưu đầu – cuối
  - Không tối ưu riêng lẻ từng thành phần
- Yêu cầu công sức của con người ít hơn
  - Không cần trích xuất đặc trưng
  - Phương pháp chung cho các cặp ngôn ngữ

# Nhược điểm của NMT?

So với dịch máy SMT:

- NMT khó giải thích
  - Khó để truy lỗi
- NMT khó kiểm soát
  - Ví dụ, không thể đưa ra các quy tắc hoặc định hướng cho việc dịch
  - Lo ngại về an toàn!



# Làm thế nào để đánh giá hệ thống MT?

---

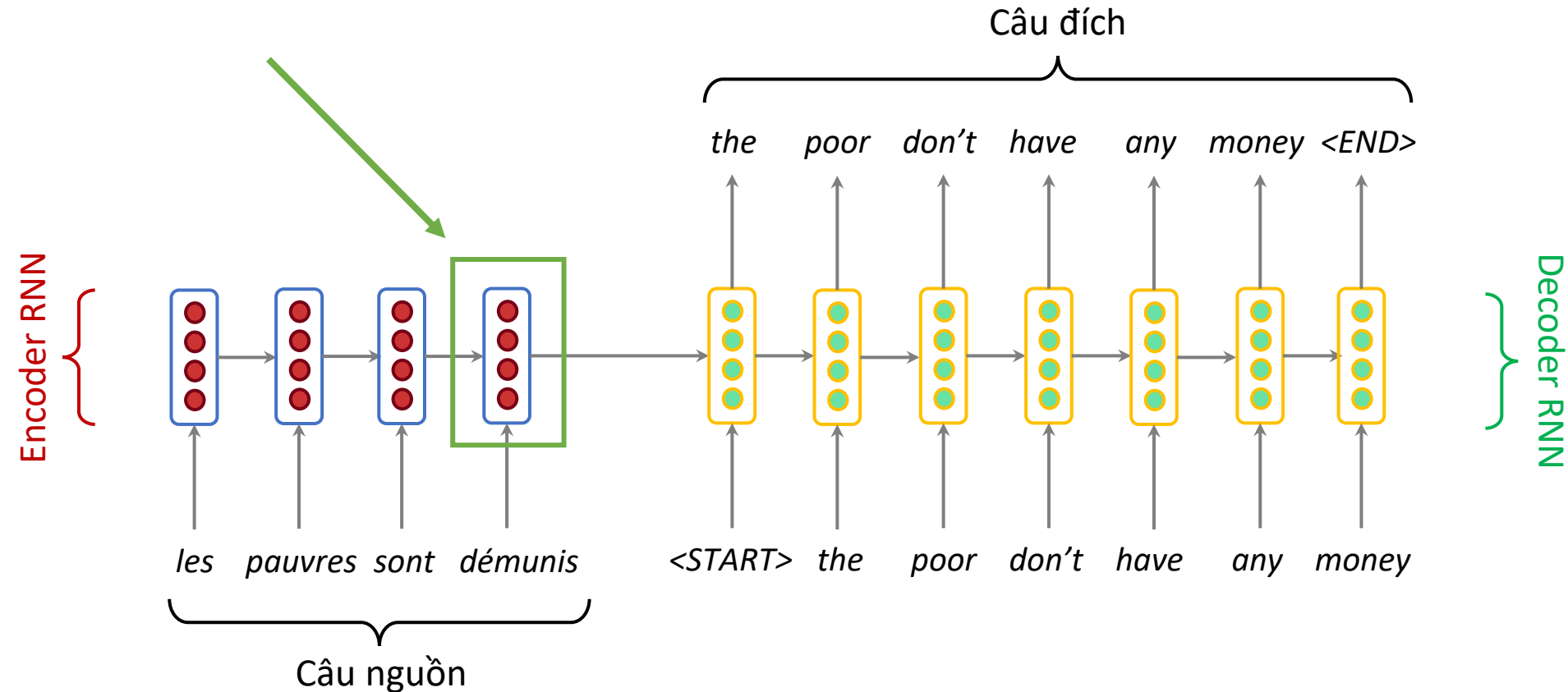
## Độ đo BLEU (Bilingual Evaluation Understudy)

- BLEU dùng để so sánh 2 văn bản
- Được áp dụng để so sánh văn bản do “máy viết”
- Độ tương tự (similarity score) được tính dựa trên:
  - $n$ -gram (thường sử dụng lên đến 3 or 4-grams)
  - Hàm phạt cho các câu dịch quá ngắn
- BLEU là một độ đo tốt nhưng không hoàn hảo
  - Có nhiều cách đánh giá câu dịch
  - Một câu dịch **tốt** có thể BLEU score **thấp** bởi vì nó có độ trùng lặp  $n$ -gram thấp so với câu người dịch 😞



# Sequence-to-sequence: Vấn đề nút cổ chai

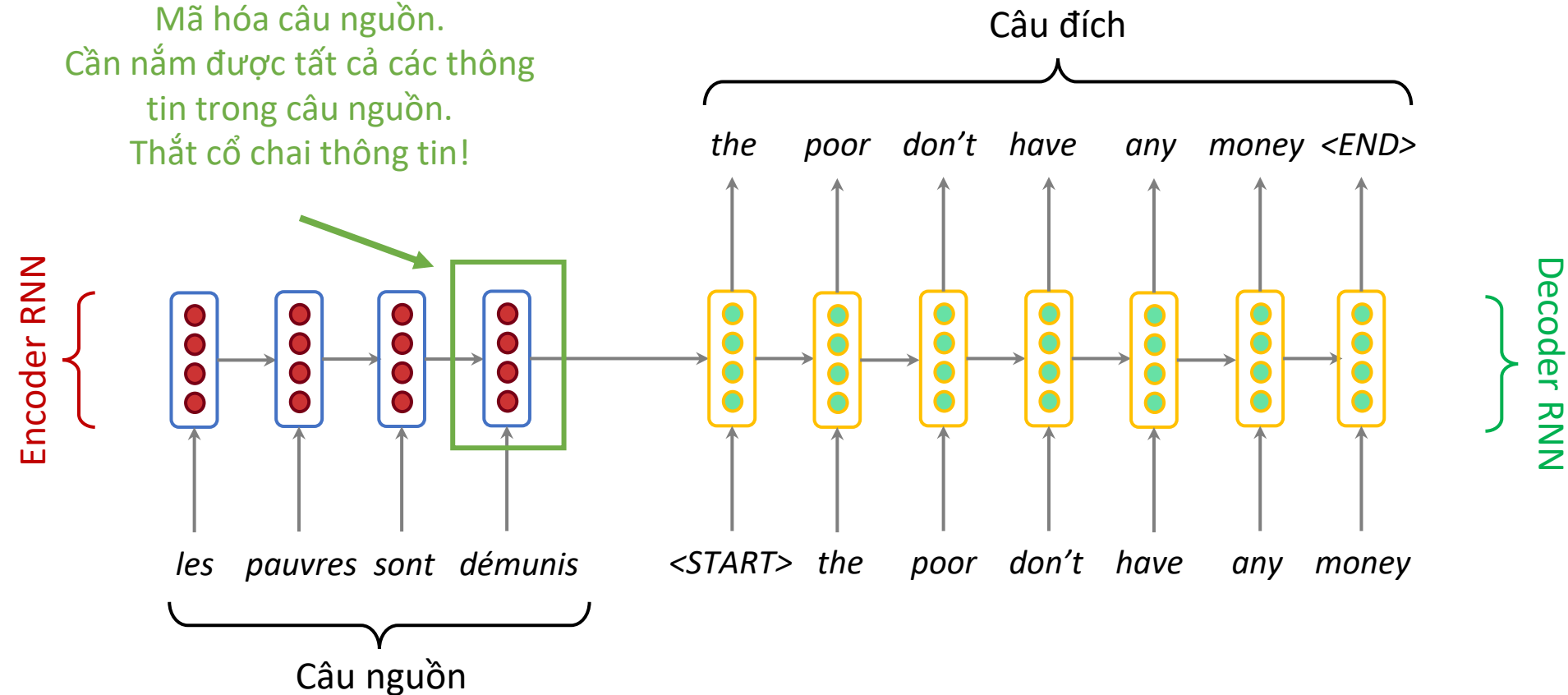
Mã hóa câu nguồn



Kiến trúc này có vấn đề gì?

# Vấn đề nút cổ chai với mô hình Seq2Seq

Mã hóa câu nguồn.  
Cần nắm được tất cả các thông tin trong câu nguồn.  
Thắt cổ chai thông tin!



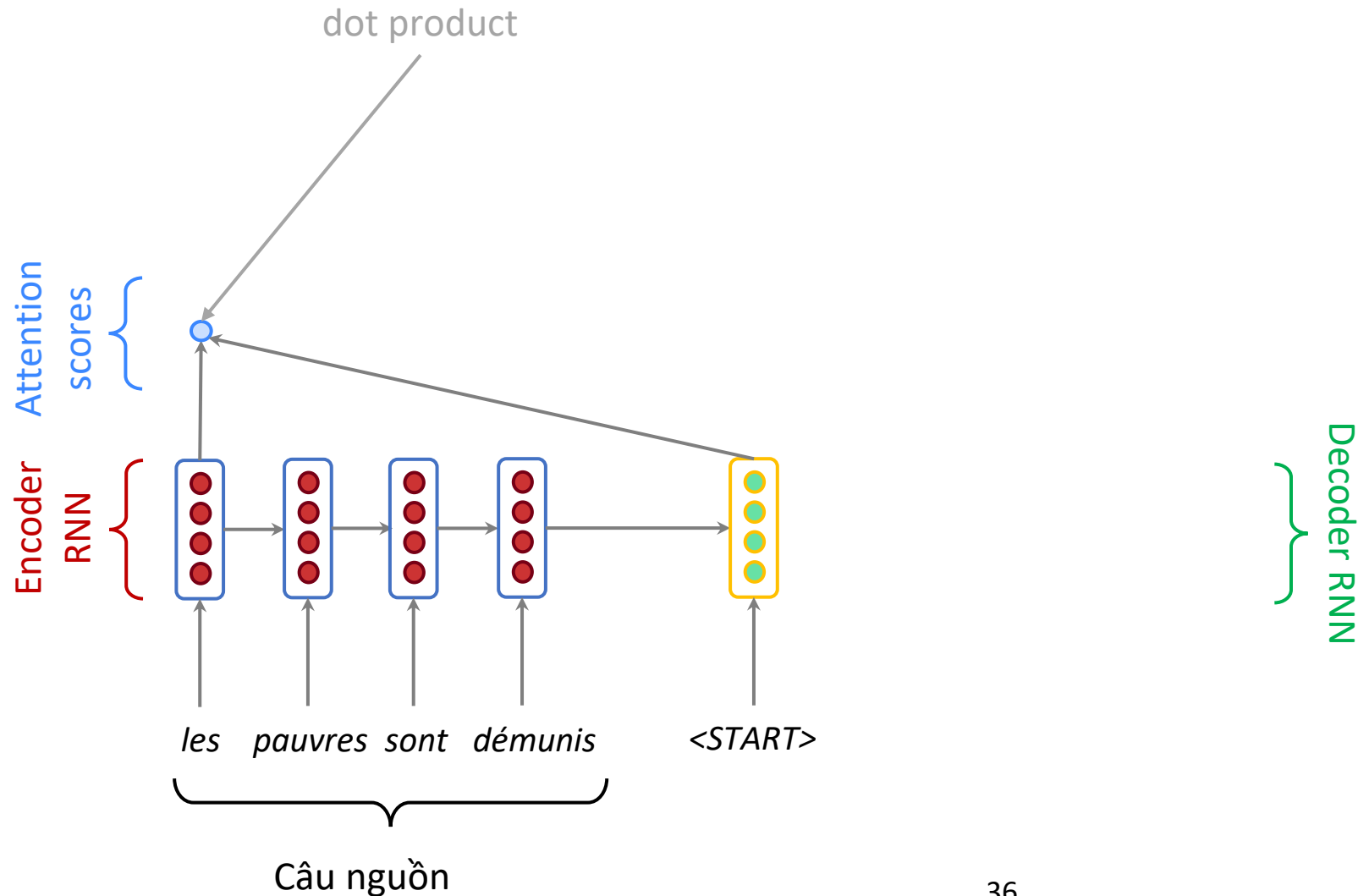
# Kỹ thuật Attention

---

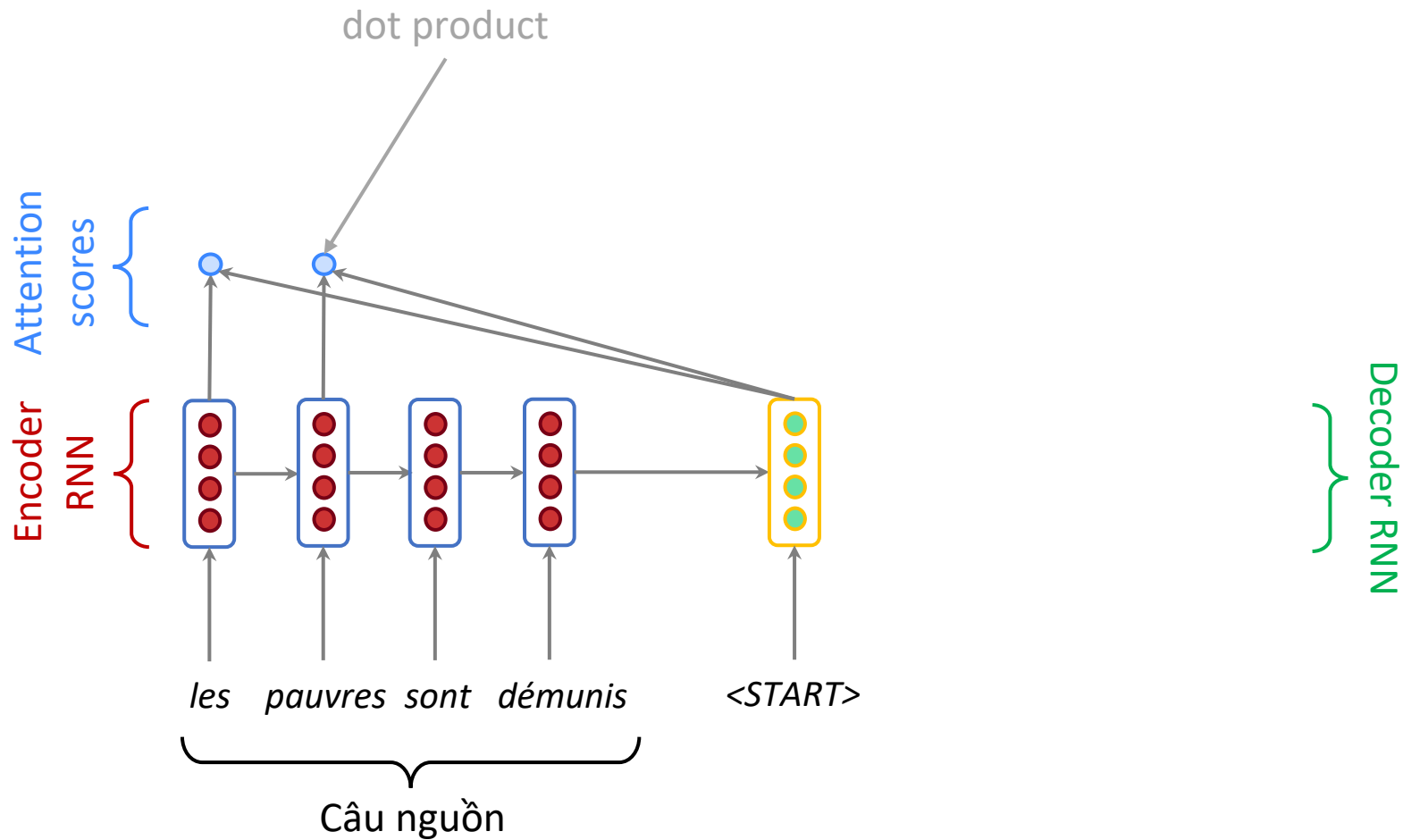
- **Attention** là một giải pháp cho thắt cổ chai.
- Ý tưởng chính: Tại mỗi bước của decoder, *tập trung vào 1 phần của câu nguồn*



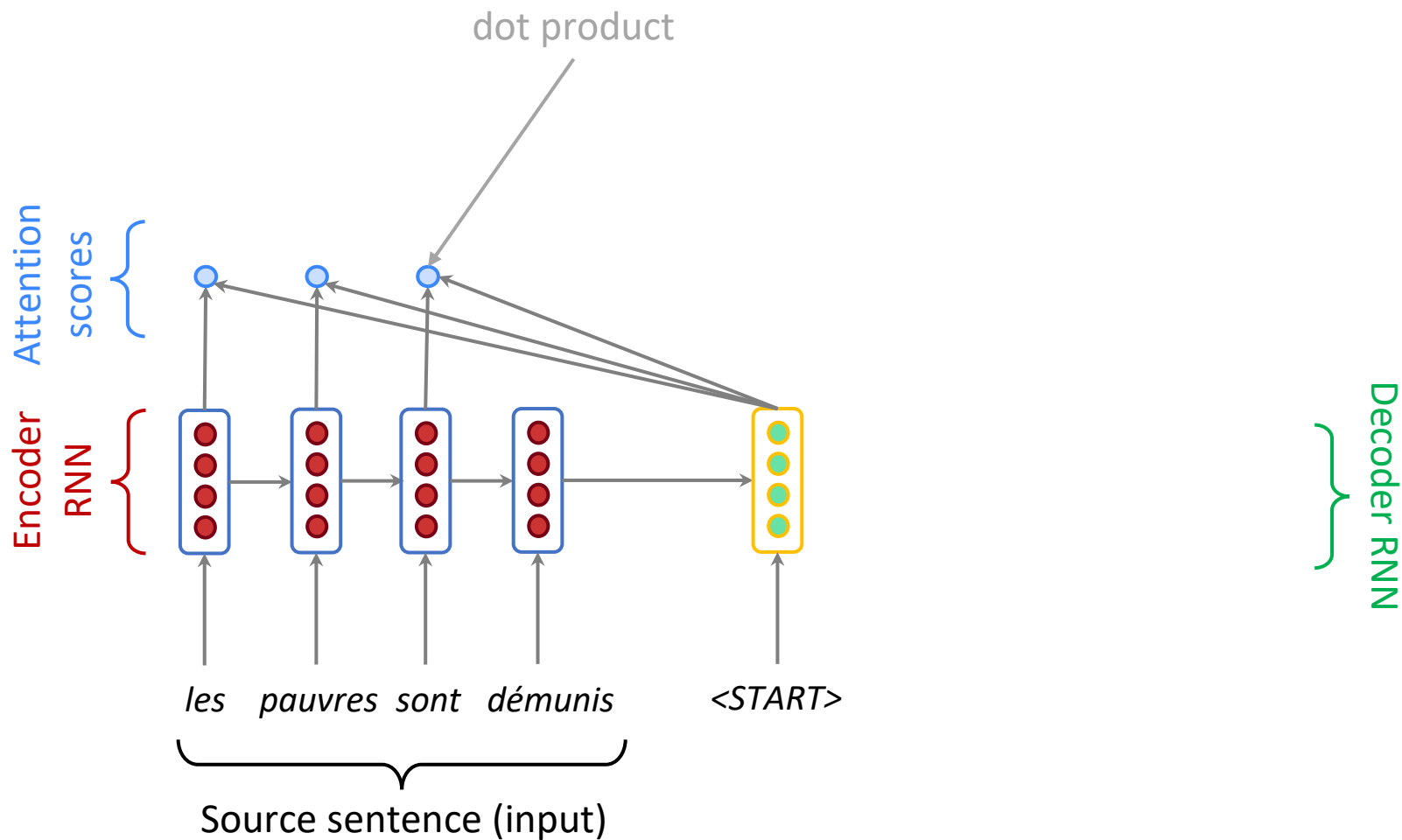
# Mô hình Encoder-Decoder với Attention



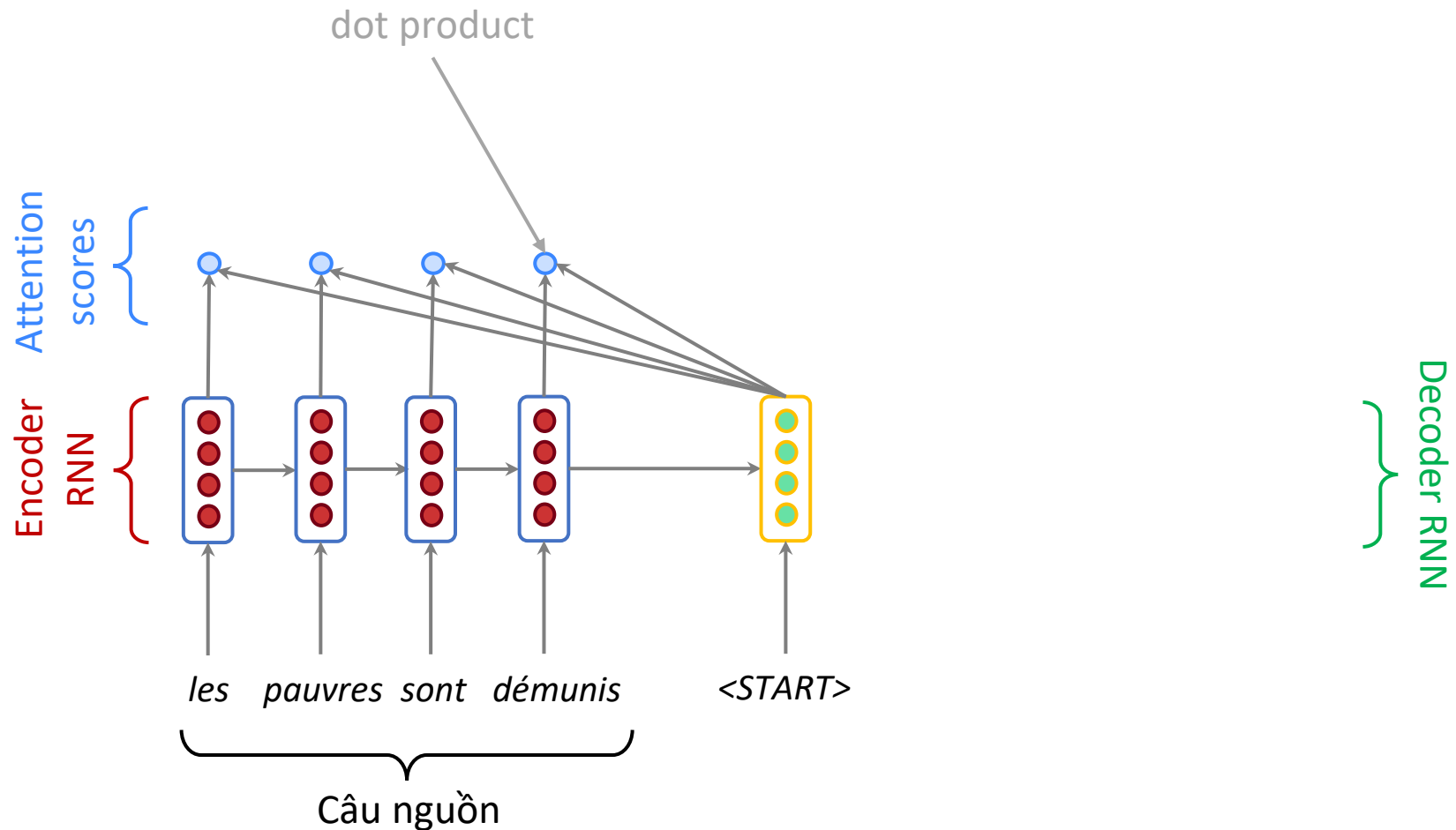
# Seq2Seq với attention



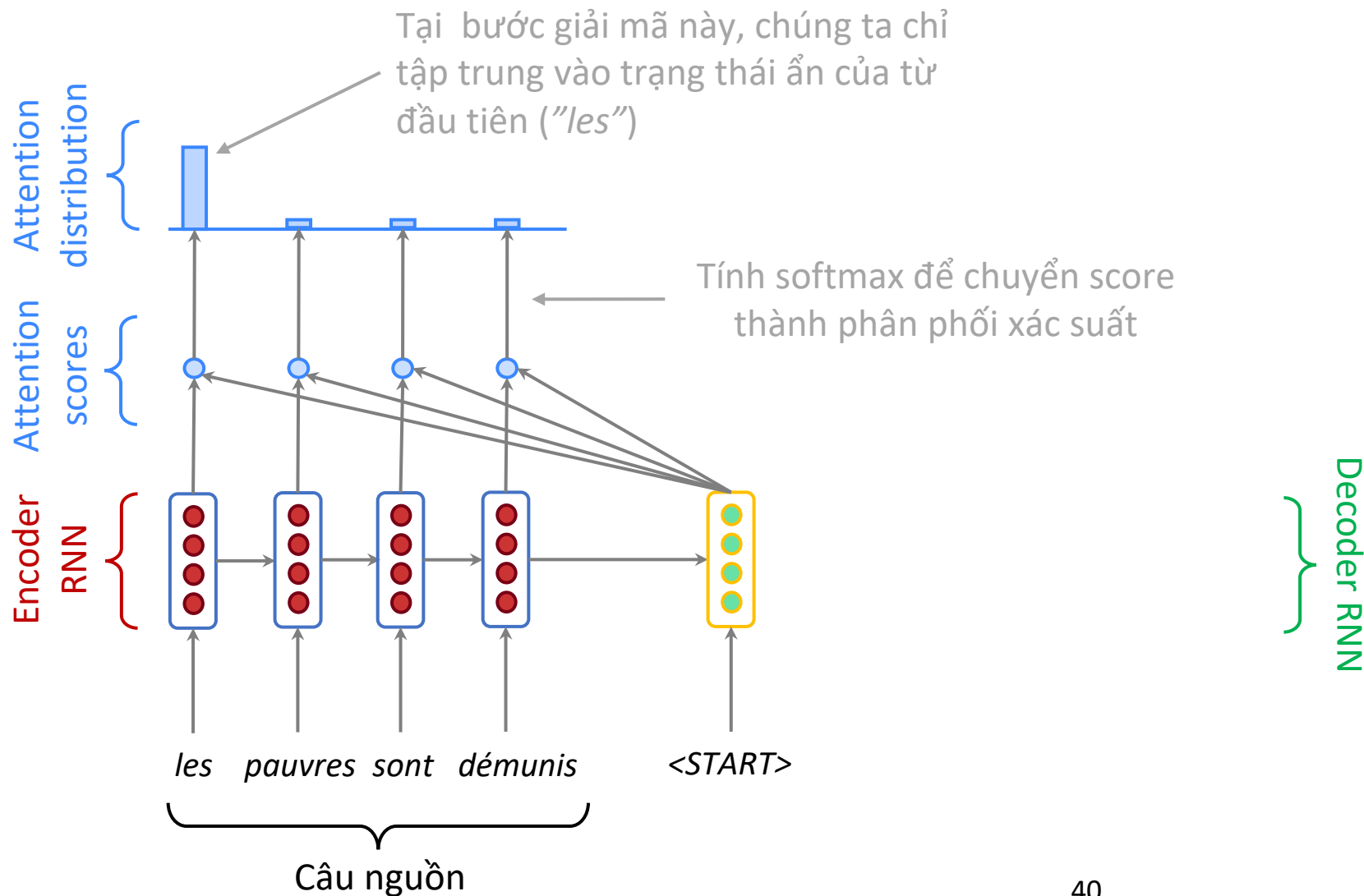
# Seq2Seq với attention



# Seq2Seq với attention

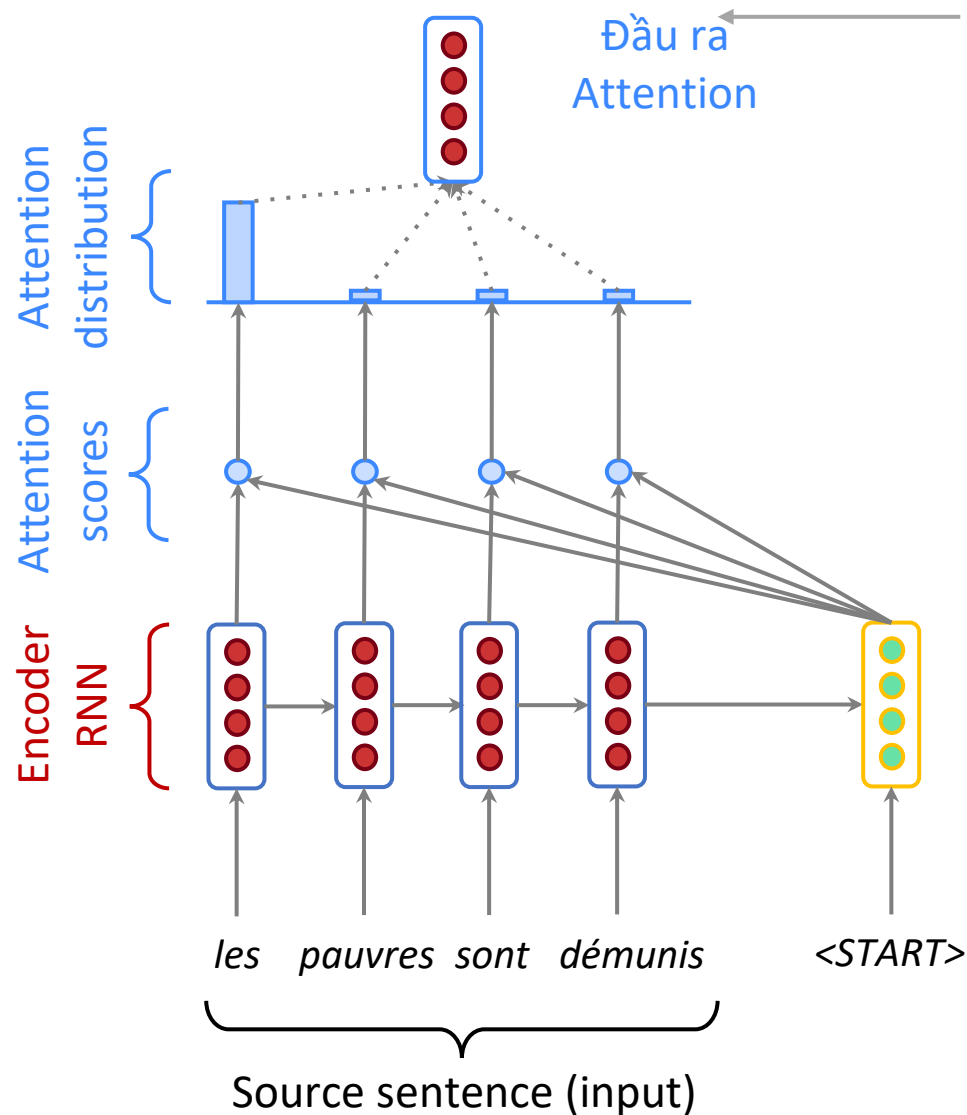


# Seq2Seq với attention





# Seq2Seq với attention

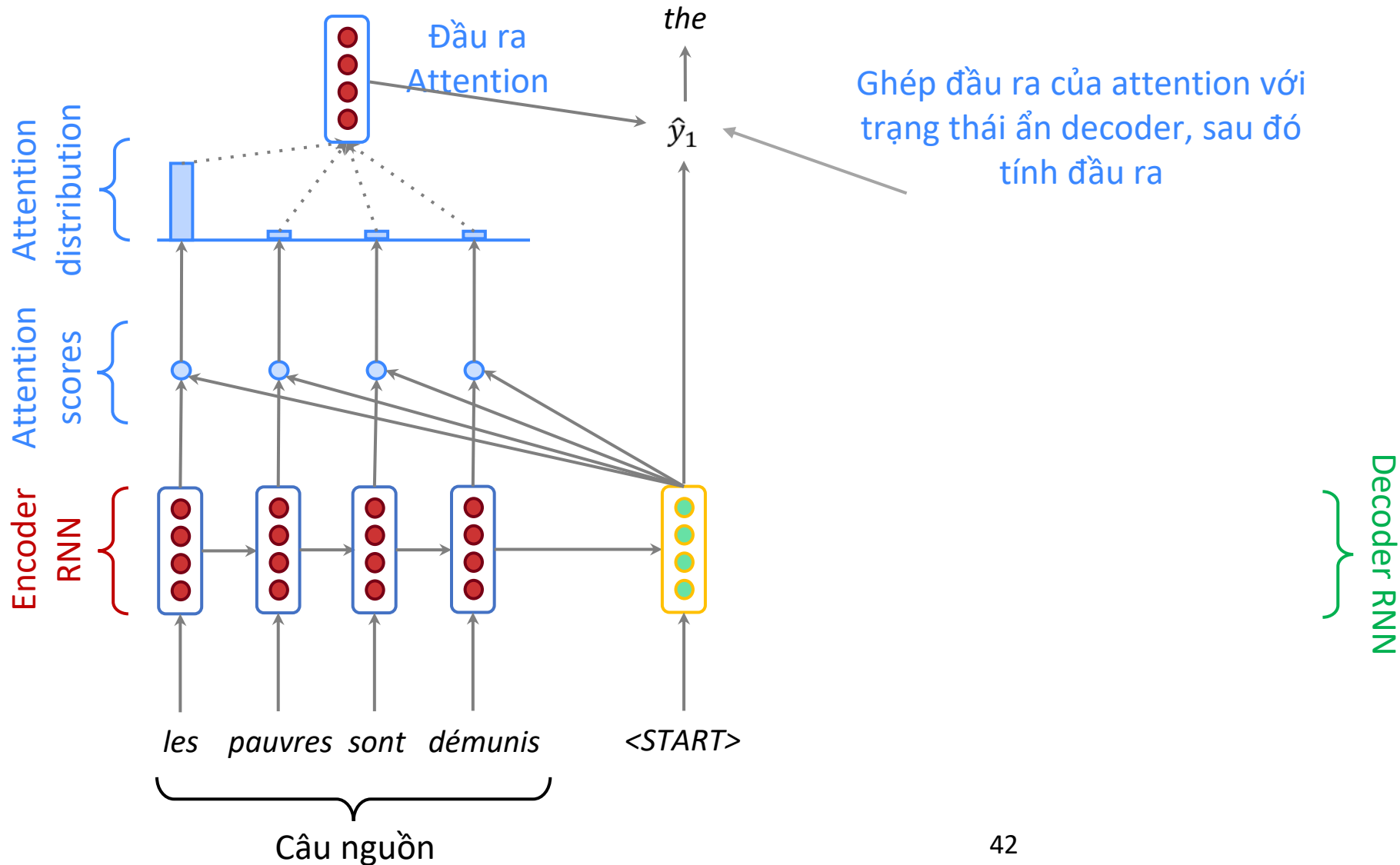


Sử dụng phân phối **attention** để tính **tổng có trọng số** của các trạng thái ẩn trong mã hóa.

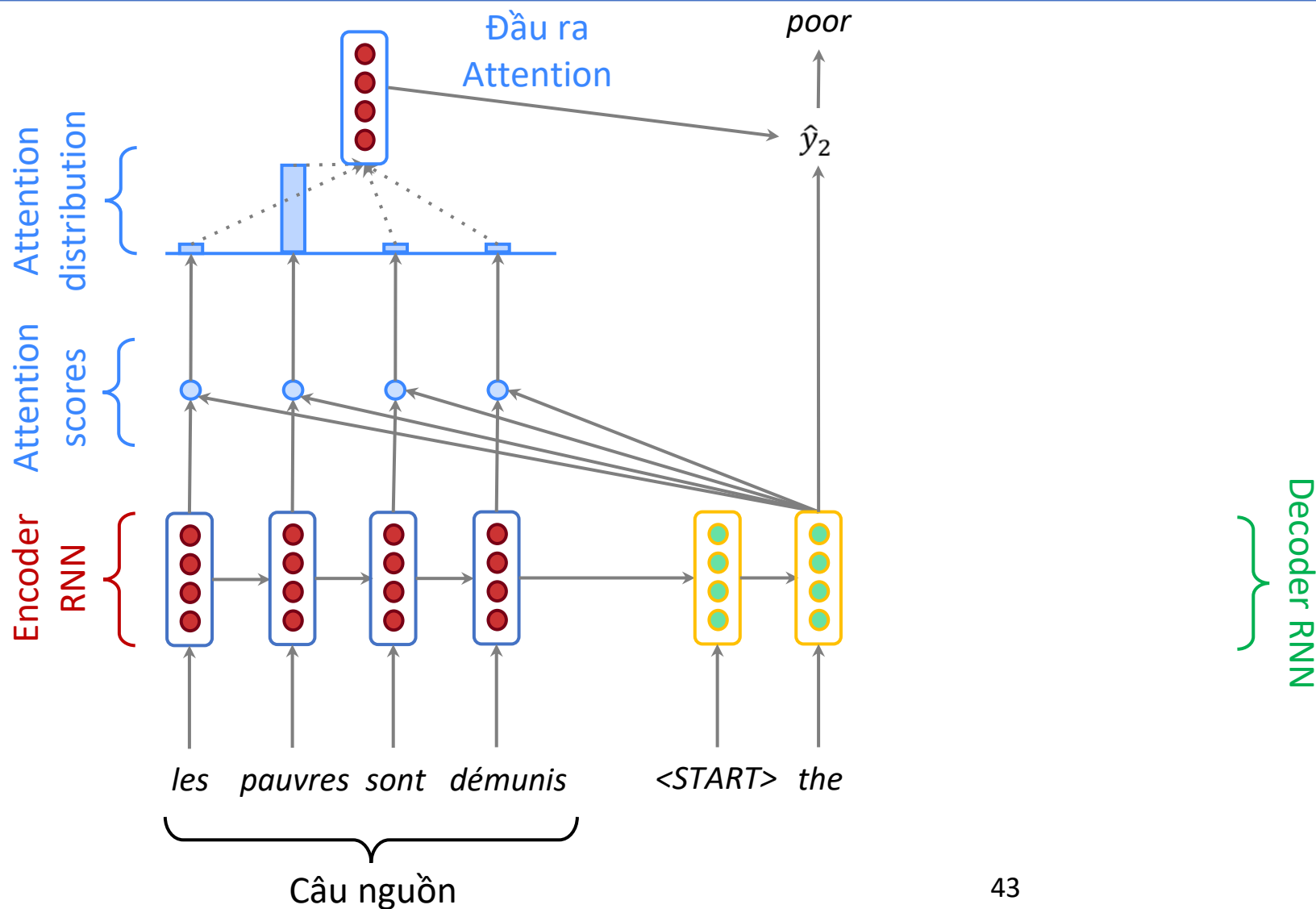
Đầu ra của **attention** phụ thuộc phần lớn vào **các trạng thái ẩn** mà nhận giá trị attention cao.

Decoder RNN

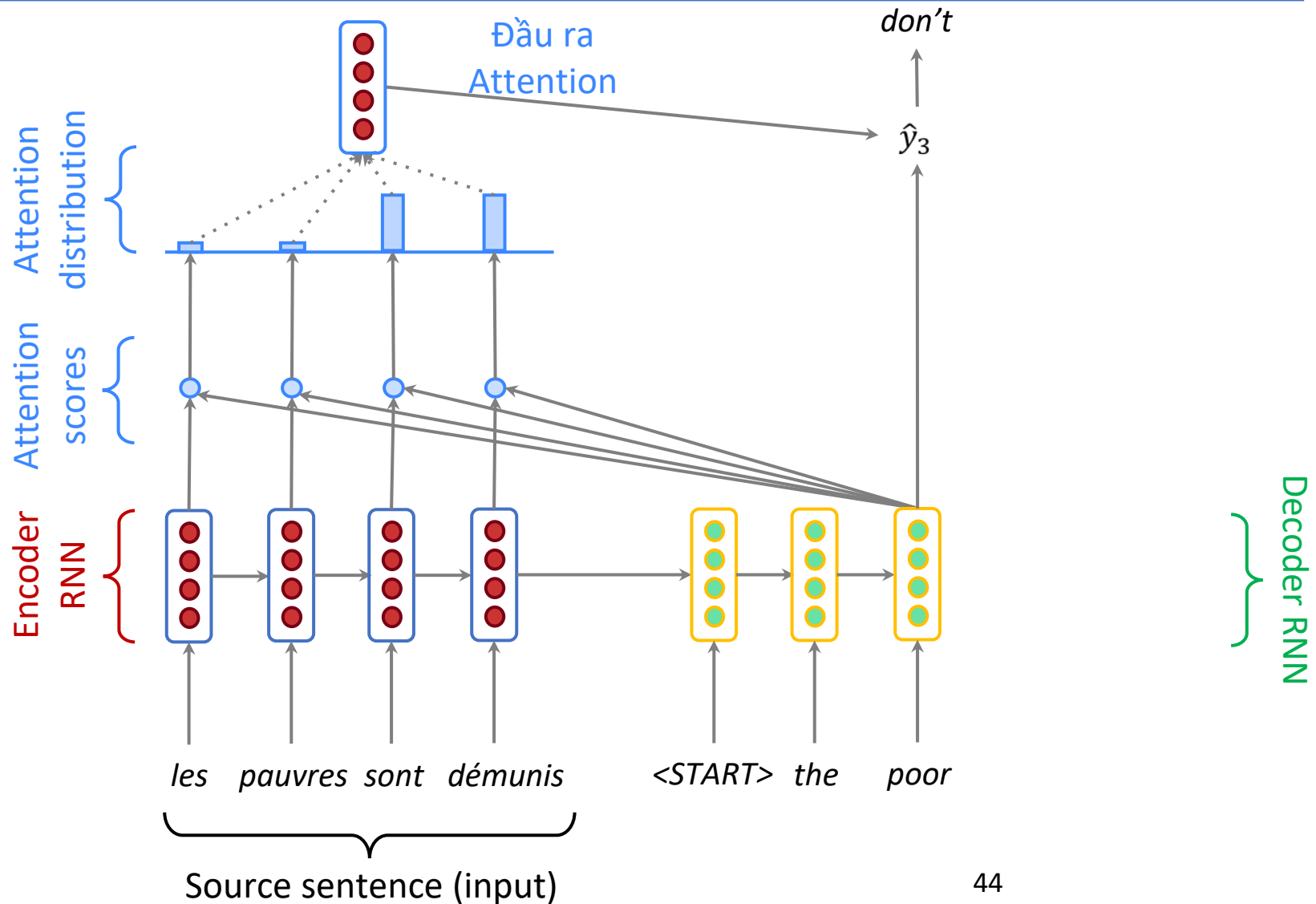
# Seq2Seq với attention



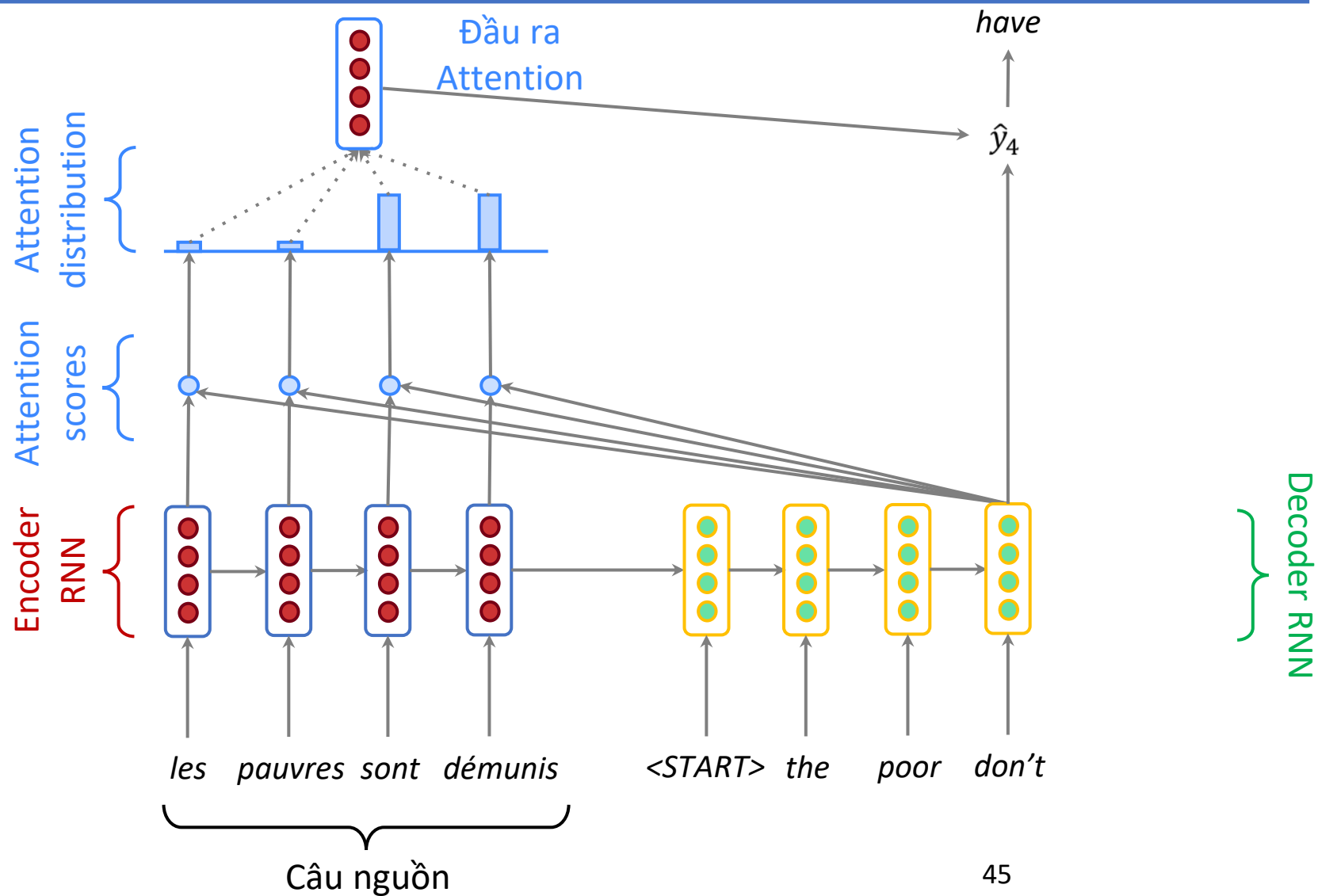
# Seq2Seq với attention



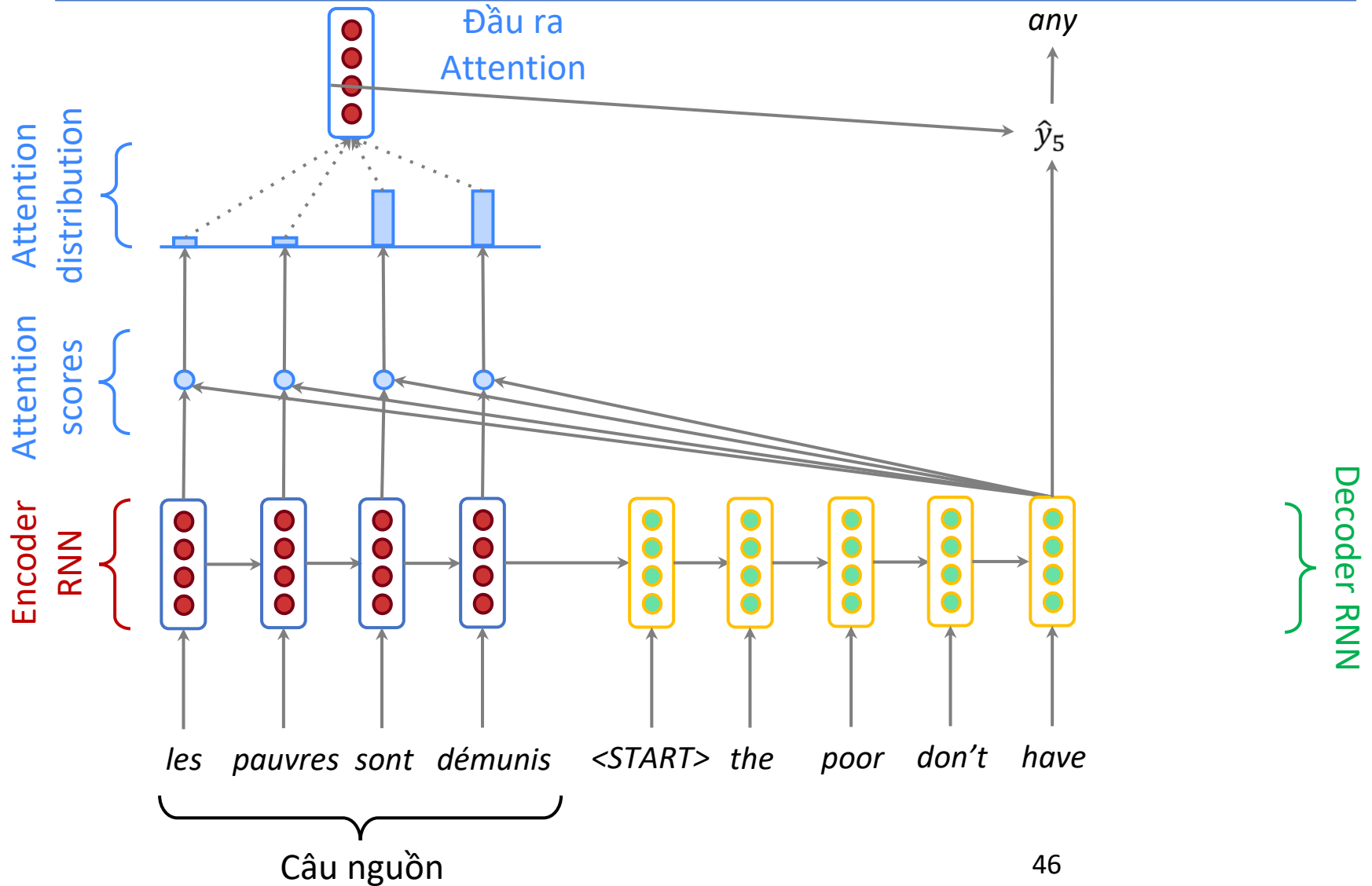
# Seq2Seq với attention



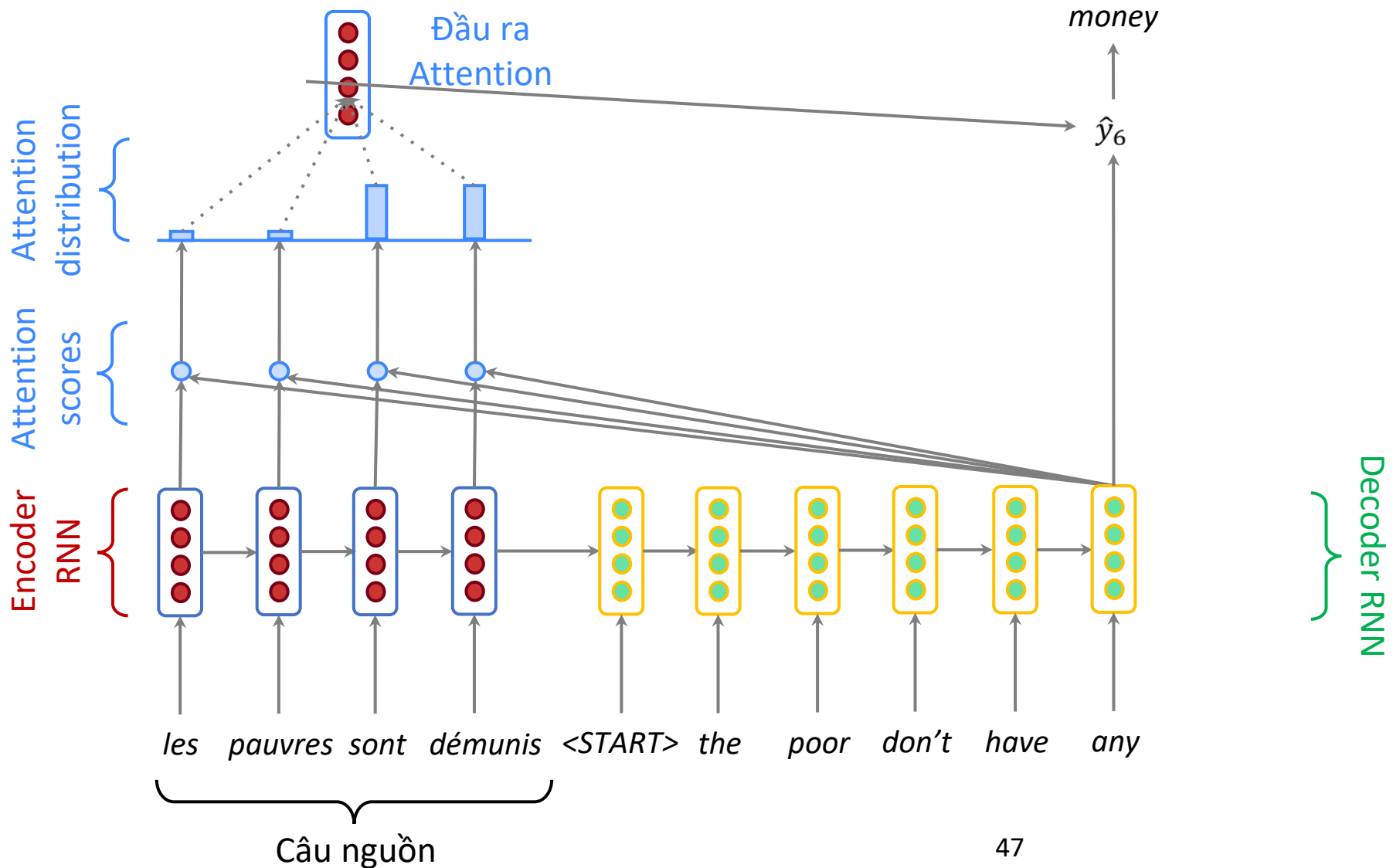
# Seq2Seq với attention



# Seq2Seq với attention



# Seq2Seq với attention



# Attention: Công thức

- Các trạng thái ẩn của encoder:  $h_1, \dots, h_N \in \mathbb{R}^h$   $s_t \in \mathbb{R}^h$
- Tại bước thời gian  $t$ , trạng thái ẩn của mã hóa  $e^t$
- Chúng ta tính điểm attention cho bước này:

$$e^t = [s_t^T h_1, \dots, s_t^T h_N] \in \mathbb{R}^N$$

- Chúng ta lấy softmax để chuyển thành phân phối xác suất  $\alpha^t$

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

- Sử dụng  $\alpha^t$  để tính tổng có trọng số của các trạng thái ẩn  $a_t$

$$a_t = \sum_{i=1}^N \alpha_i^t h_i \in \mathbb{R}^h$$

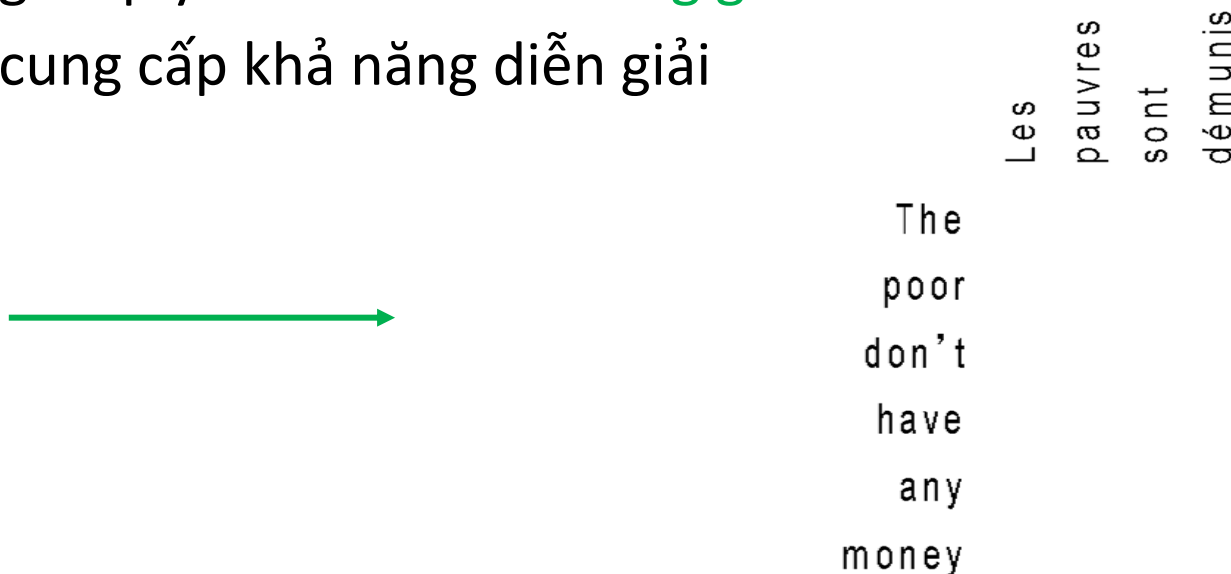
- Nối đầu ra của vector attention  $a_t$  với trạng thái ẩn của encoder  $s_t$  và thực hiện xử lý như mô hình seq2seq thông thường

$$[a_t; s_t] \in \mathbb{R}^{2h}$$



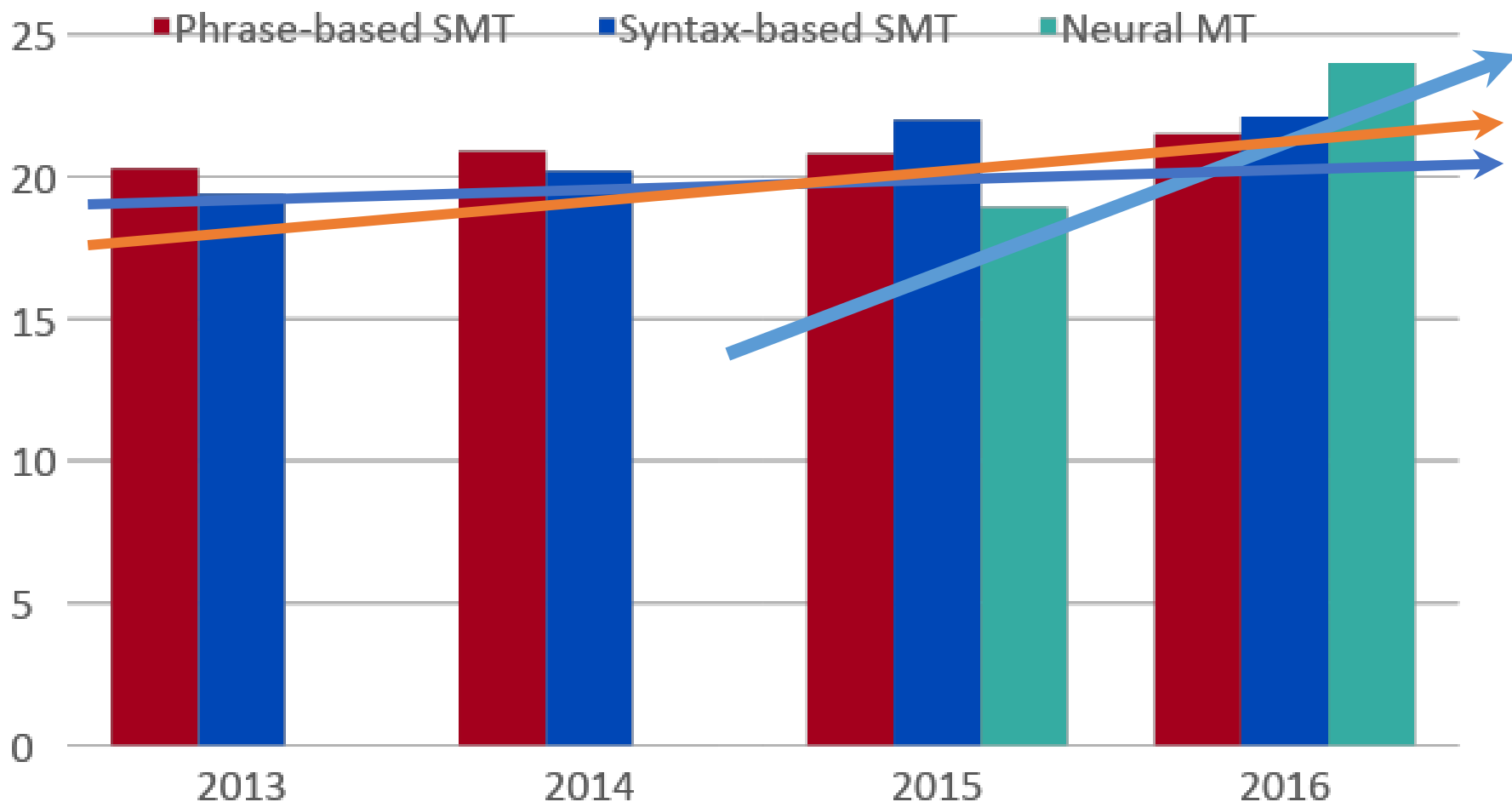
# Attention tuyệt vời

- Attention cải thiện đáng kể hiệu năng của **NMT**
  - Attention rất hiệu quả cho phép decoder tập trung trên các phần nào đó của câu nguồn
- Attention **giải quyết vấn đề thắt cổ chai**
- Attention giải quyết vấn đề **vanishing gradient**
- Attention cung cấp khả năng diễn giải



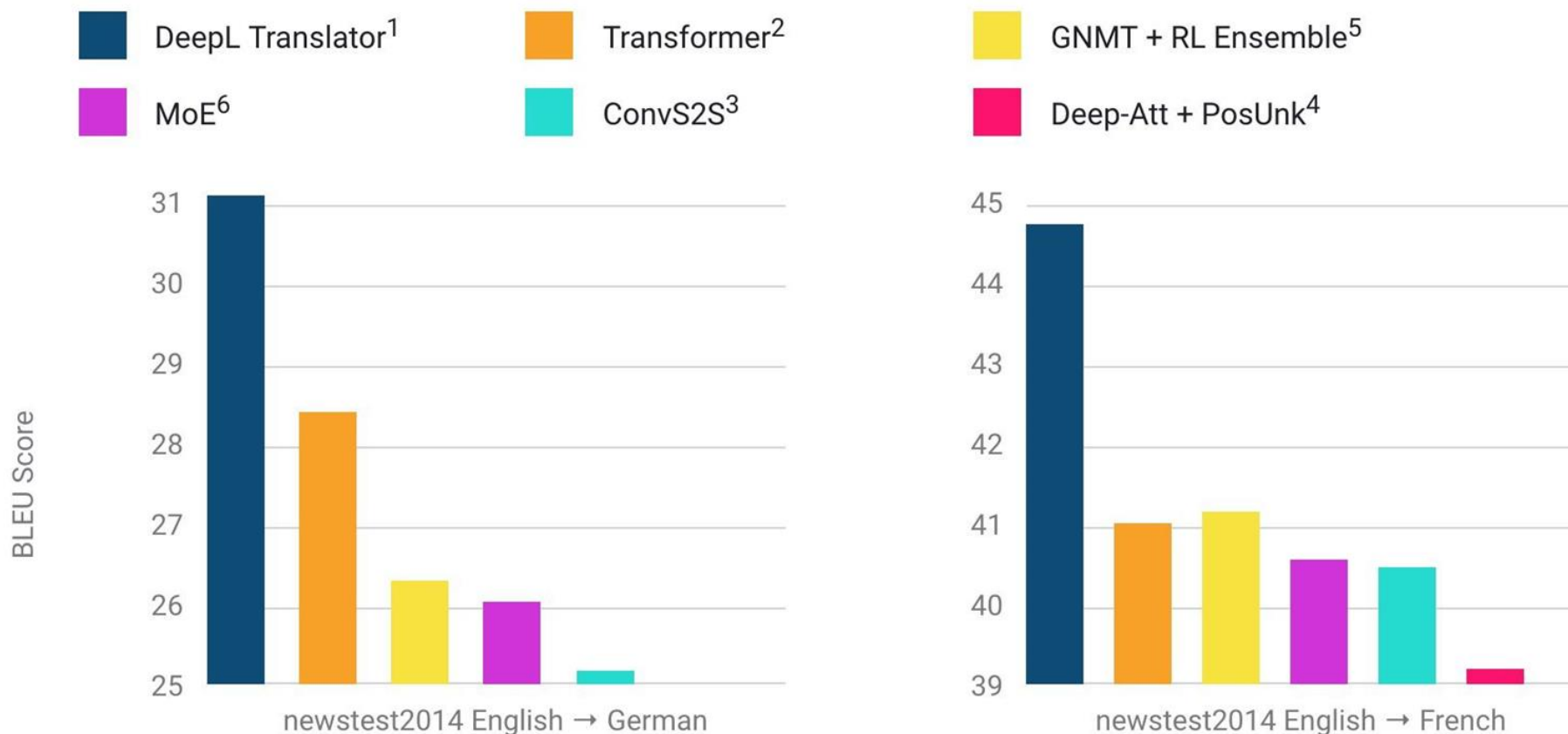
# Tiến triển của hệ thống MT theo thời gian

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



Source: [http://www.meta-net.eu/events/meta-forum-2016/slides/09\\_sennrich.pdf](http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf)

# Vấn đề dữ liệu



Source: DeepL's [press release](#) (Aug 2017)

# NMT: Câu chuyện thành công lớn nhất của NLP Deep Learning

---

Dịch máy mạng nơ ron bắt đầu được nghiên cứu từ **2014**, trở thành **Phương pháp hàng đầu 2016**

- **2014:** Bài báo seq2seq được công bố
- **2016:** Google Translate chuyển từ SMT sang NMT
- **Đáng kinh ngạc!**
  - **SMT** hệ thống, xây dựng bởi hàng trăm kỹ sư qua nhiều năm, bị vượt qua bởi hệ thống NMT huấn luyện bởi một số kỹ sư trong vài tháng

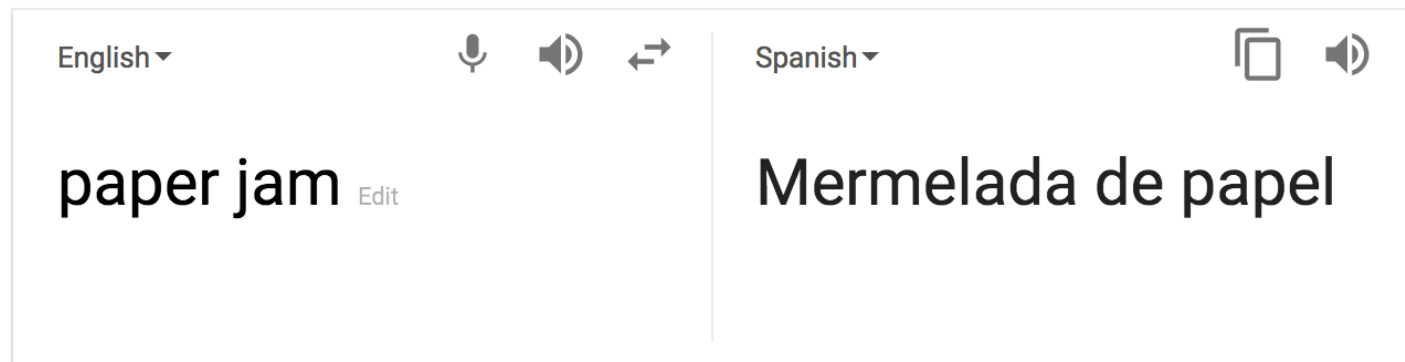
# Vậy là MT đã được giải quyết chưa?

---

- **Chưa!**
- Vẫn còn nhiều khó khăn cần giải quyết:
  - Các từ ngoài từ điển (OOV)
  - Khác lĩnh vực giữa dữ liệu huấn luyện và kiểm tra
  - Duy trì ngữ cảnh đối với văn bản dài
  - Các cặp ngôn ngữ ít dữ liệu

# Vậy là MT đã được giải quyết chưa?

- Chưa!



[Open in Google Translate](#)

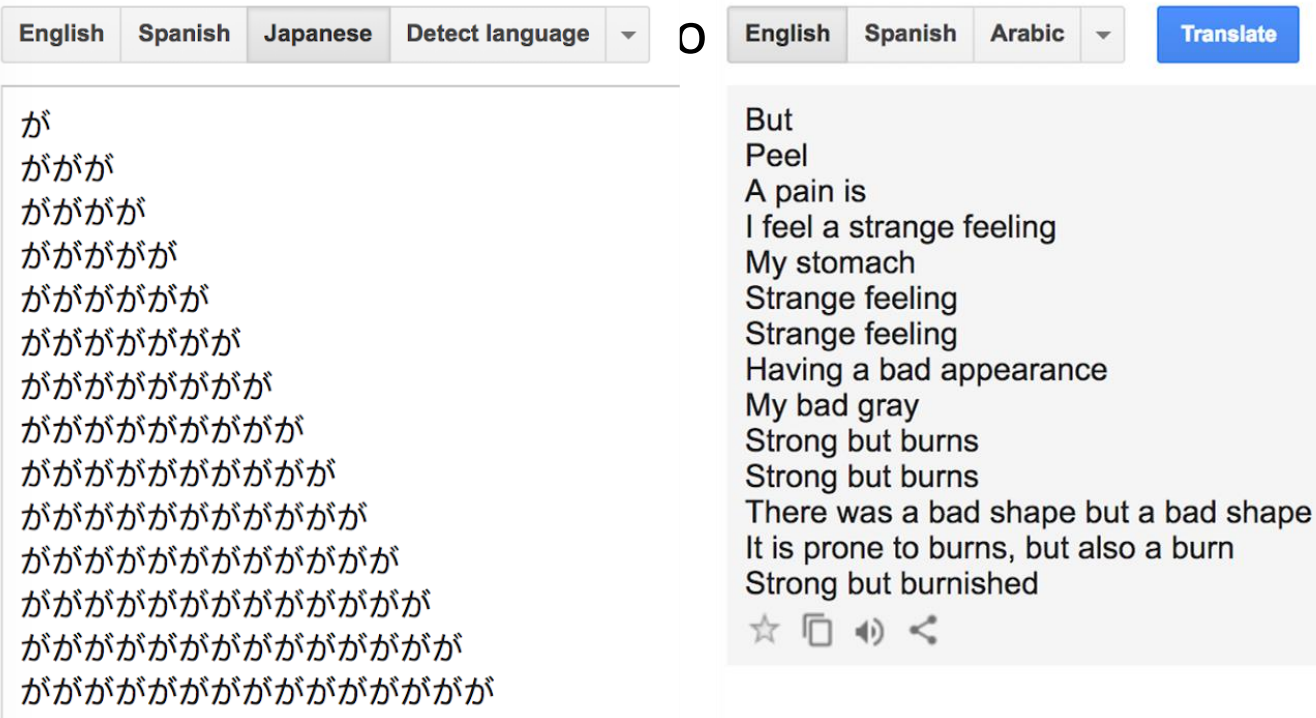
[Feedback](#)



?

# Vậy là MT đã được giải quyết chưa?

- Nope!

- 

The screenshot shows a Google Translate interface. On the left, the input is a sequence of 14 'か' characters. The right side shows the output in English, which is a list of phrases that are phonetic interpretations of the sound 'ka'.

English	Spanish	Japanese	Detect language
English	Spanish	Japanese	Detect language

か  
かかか  
かかかか  
かかかかか  
かかかかかか  
かかかかかかか  
かかかかかかかか  
かかかかかかかかか  
かかかかかかかかかか  
かかかかかかかかかかか  
かかかかかかかかかかかか  
かかかかかかかかかかかか  
かかかかかかかかかかかかか  
かかかかかかかかかかかかか

But  
Peel  
A pain is  
I feel a strange feeling  
My stomach  
Strange feeling  
Strange feeling  
Having a bad appearance  
My bad gray  
Strong but burns  
Strong but burns  
There was a bad shape but a bad shape  
It is prone to burns, but also a burn  
Strong but burnished

Source: <http://language-log.ldc.upenn.edu/nll/?p=35120#more-35120>

# Seq2seq là rất linh hoạt!

---

- Mô hình Seq2Seq là hữu ích không chỉ cho *MT*
- Rất nhiều bài toán NLP tasks có thể sử dụng Seq2Seq:
  - Tóm tắt văn bản (văn bản dài → văn bản ngắn)
  - Hội thoại (Câu trước → Câu sau)
  - Phân tích (Văn bản đầu vào → chuỗi phân tích đầu ra)
  - Sinh mã nguồn (ngôn ngữ tự nhiên → Mã Python)



# Kết luận

---

- Từ 2014, Dịch máy mạng nơ ron thay thế dịch máy thống kê
- Seq2Seq là một kiến trúc cho dịch máy (sử dụng 2 RNNs)
- Attention là một cách để tập trung vào các phần cụ thể của đầu vào
  - Nâng cao mô hình Seq2Seq rất lớn

