VietAI

Bài 14: Machine Translation and Sequence to Sequence model

Tuần 7B

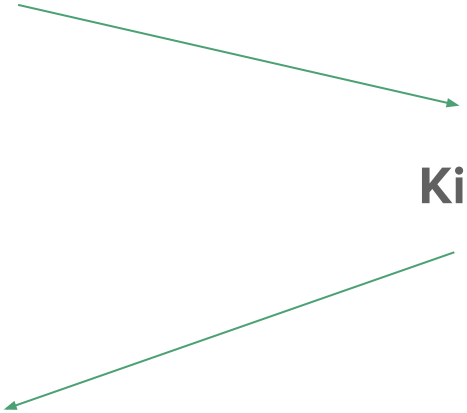
03-10-2019

Nội dung chính

1. Dịch máy (machine translation)
    a. Pre-Neural Machine Translation
    b. Neural Machine Translation
2. Kiến trúc Seq2Seq
3. Một số ứng dụng thực tế của mô hình Seq2Seq
    a. Dịch máy (Neural machine translation)
    b. Nhận diện giọng nói (Speech recognition)
    c. Image captioning

Nội dung chính

Task: Dịch máy (machine translation)



Kiến trúc: Seq2Seq

Concept: Conditional Language model

VietAI

1a. Pre-Neural Machine Translation

Pre-Neural Machine Translation

Machine Translation (MT) is the task of translating a sentence *x* from one language (the source language) to a sentence *y* in another language (the target language).

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

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

- Rousseau

1950s: Early Machine Translation

Machine Translation research began in the early 1950s.

- Russian → English
  (motivated by the Cold War!)

1 minute video showing 1954 MT:
https://youtu.be/K-HfpsHPmvw

- Systems were mostly rule-based, using a bilingual dictionary to map Russian words to their English counterparts

1990s-2010s: Statistical Machine Translation

- Core idea: Learn a probabilistic model from data
- Suppose we're translating French → English.
- We want to find best English sentence *y*, given French sentence *x*

$$\text{argmax}_y P(y|x)$$

1990s-2010s: Statistical Machine Translation

- Core idea: Learn a probabilistic model from data
- Suppose we're translating French → English.
- We want to find best English sentence *y*, given French sentence *x*

$$\text{argmax}_y P(y|x)$$

- Use Bayes Rule to break this down into two components to be learnt separately:

$$= \text{argmax}_y P(x|y)P(y)$$

1990s-2010s: Statistical Machine Translation

- Core idea: Learn a probabilistic model from data
- Suppose we're translating French → English.
- We want to find best English sentence *y*, given French sentence *x*

$$\text{argmax}_y P(y|x)$$

- Use Bayes Rule to break this down into two components to be learnt separately:

$$= \text{argmax}_y \underbrace{P(x|y)}_{\text{Translation Model}} \underbrace{P(y)}_{\text{Language Model}}$$

Translation Model

Models how words and phrases should be translated (*fidelity*). Learnt from parallel data.

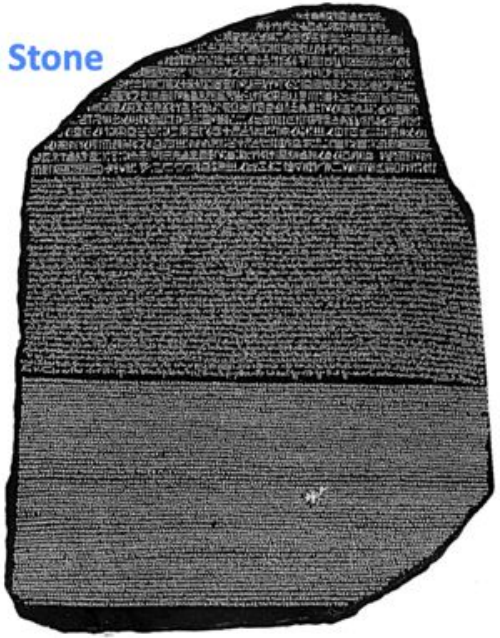
Language Model

Models how to write good English (*fluency*). Learnt from monolingual data.

1990s-2010s: Statistical Machine Translation

- Question: How to learn translation model $P(x|y)$?
- First, need large amount of parallel data
  (e.g. pairs of human-translated French/English sentences)

The Rosetta Stone



Ancient Egyptian

Demotic

Ancient Greek

Learning alignment for SMT

- Question: How to learn translation model $P(x|y)$ from the parallel corpus?

Learning alignment for SMT

- Question: How to learn translation model $P(x|y)$ from the parallel corpus?

- Break it down further: we actually want to consider

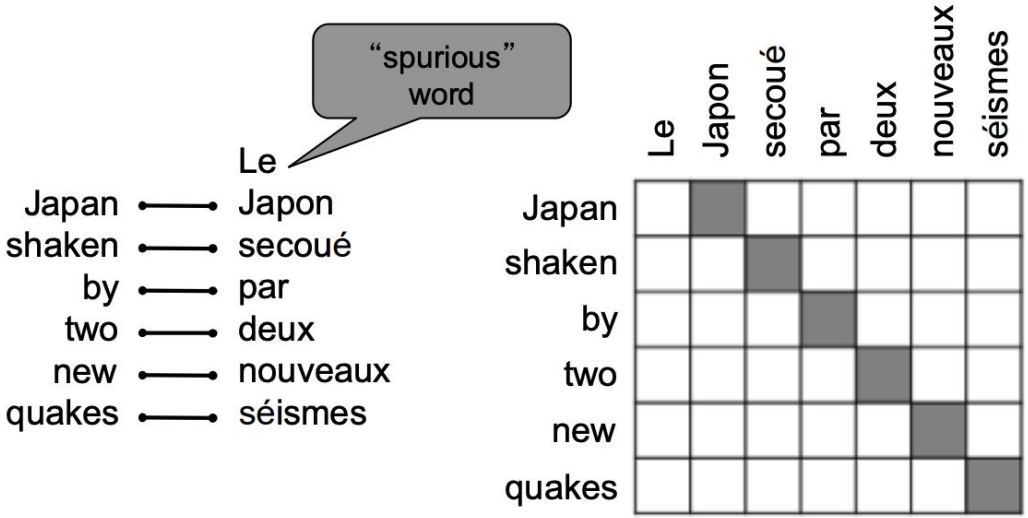
$$P(x, a|y)$$

where *a* is the alignment, i.e. word-level correspondence between French sentence *x* and English sentence *y*

What is alignment?

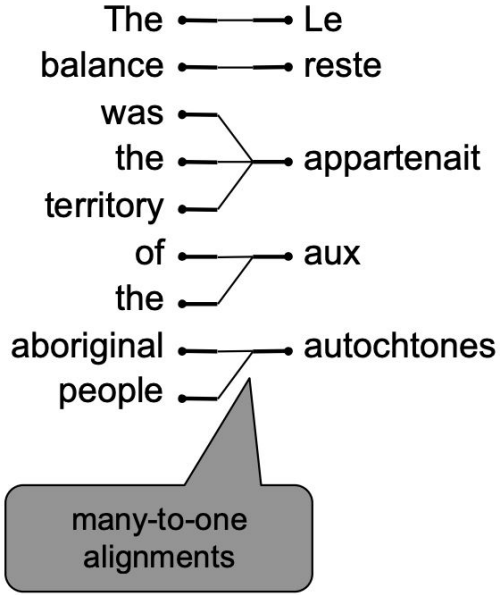
Alignment is the correspondence between particular words in the translated sentence pair.

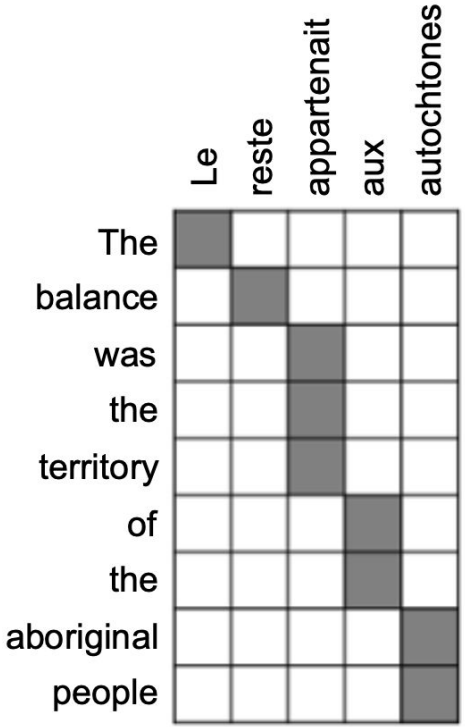
- Note: Some words have no counterpart



Alignment is complex

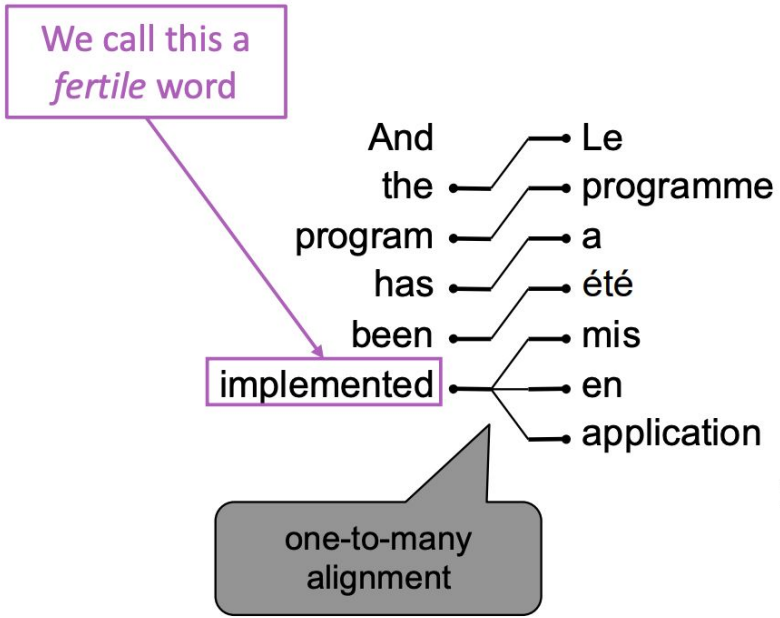
Alignment can be many-to-one

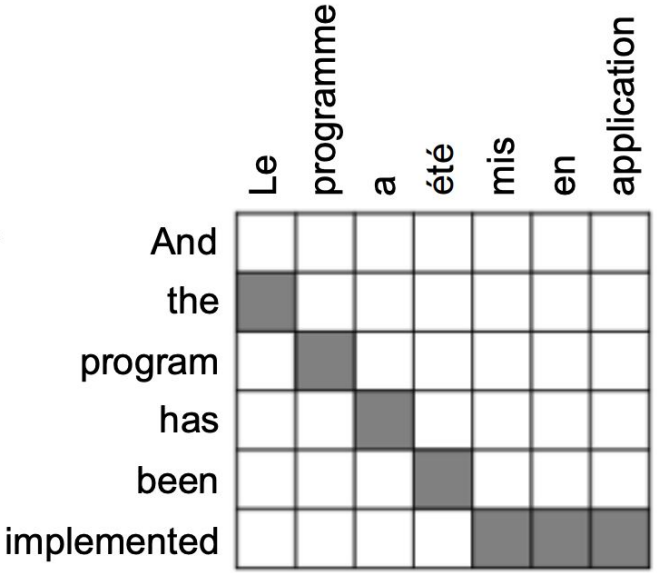




Alignment is complex

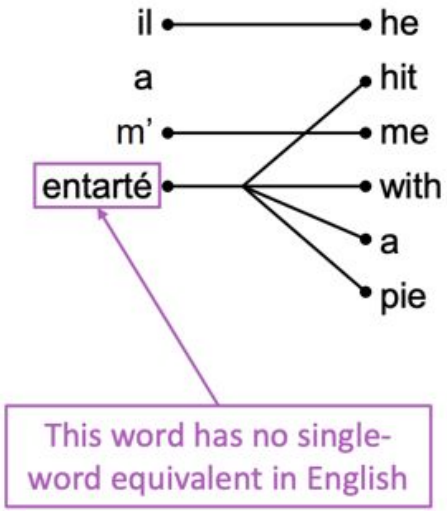
Alignment can be one-to-many

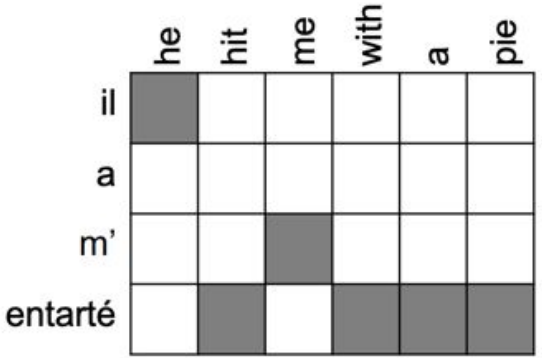




Alignment is complex

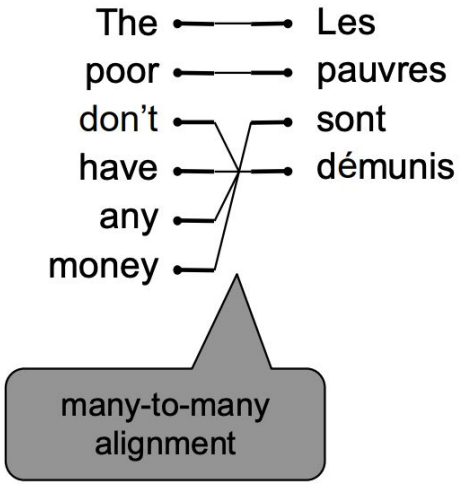
Some words are very fertile!

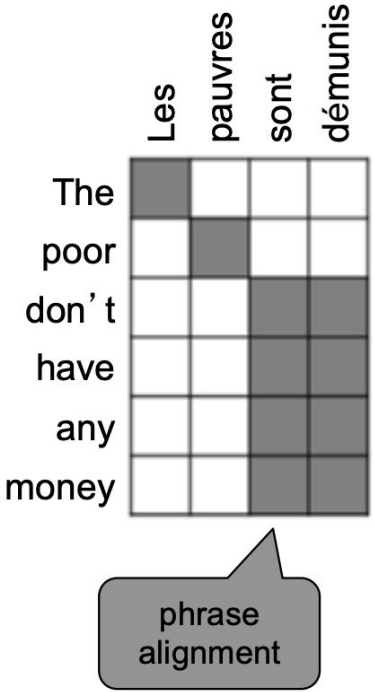




Alignment is complex

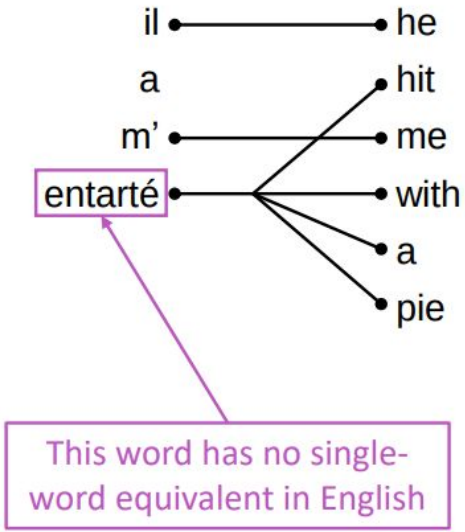
Alignment can be many-to-many (phrase-level)

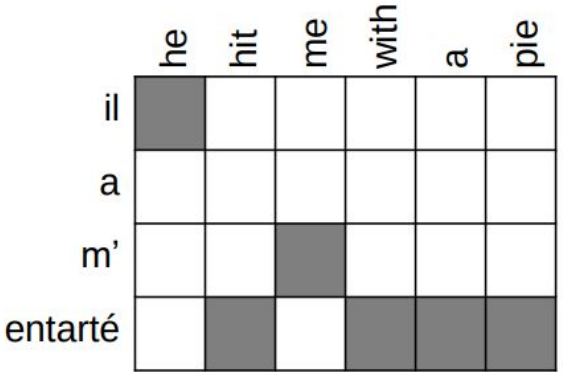




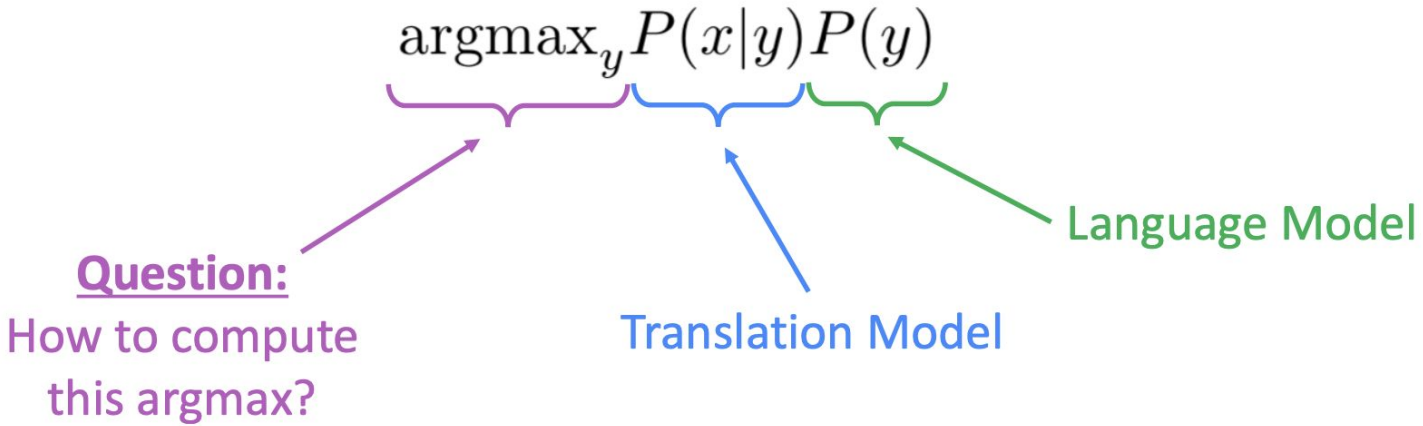
Alignment is complex

Some words are very fertile!



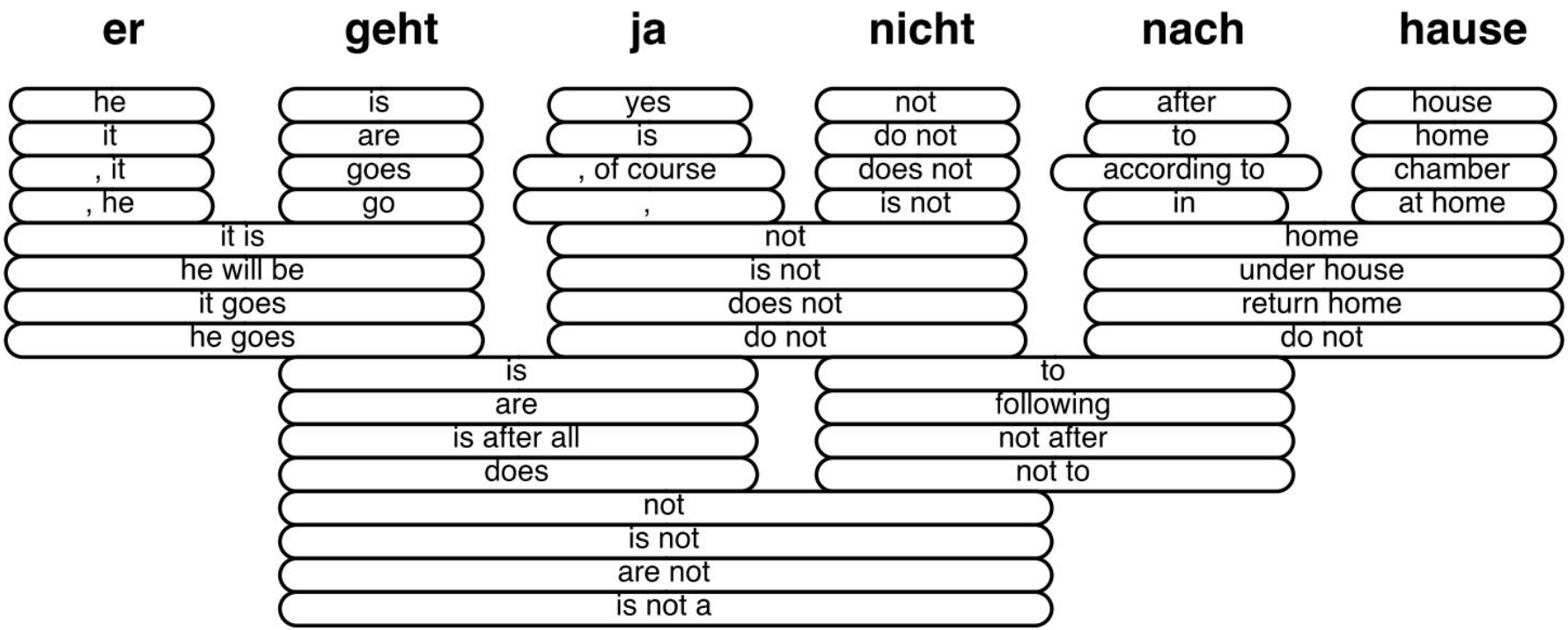


Decoding for SMT



- We could enumerate every possible *y* and calculate the probability? → Too expensive!
- **Answer:** Use a heuristic search algorithm to search for the best translation, discarding hypotheses that are too low-probability
- This process is called *decoding*

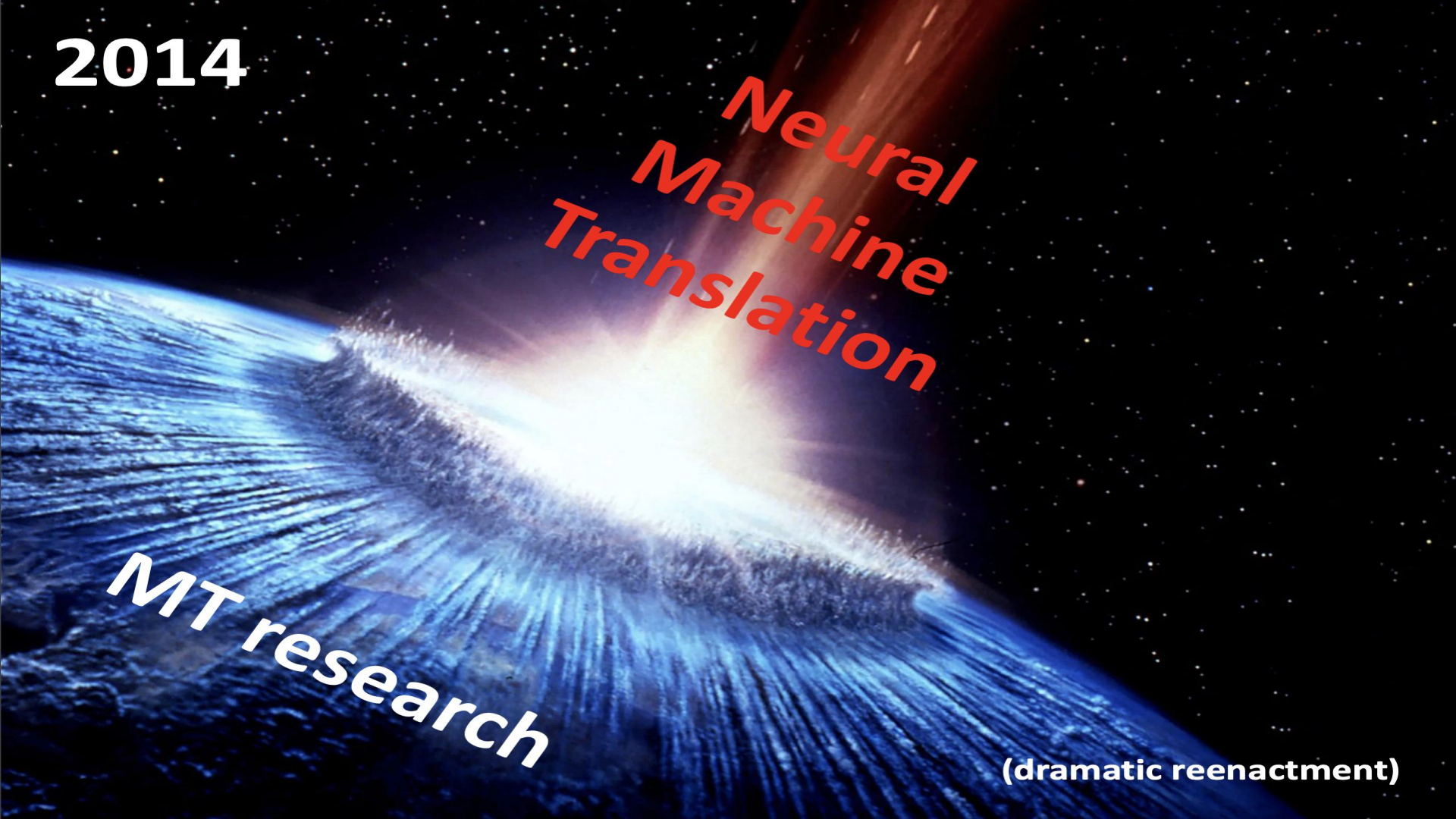
Decoding for SMT



1990s-2010s: Statistical Machine Translation

VietAI

- SMT was a huge research field
- The best systems were extremely complex
  - Hundreds of important details we haven't mentioned here
  - Systems had many separately-designed subcomponents
  - Lots of feature engineering
    - Need to design features to capture particular language phenomena
  - Require compiling and maintaining extra resources
    - Like tables of equivalent phrases
  - Lots of human effort to maintain
    - Repeated effort for each language pair!


2014
Neural
Machine
Translation
MT research
(dramatic reenactment)

VietAI

1b. Neural Machine Translation

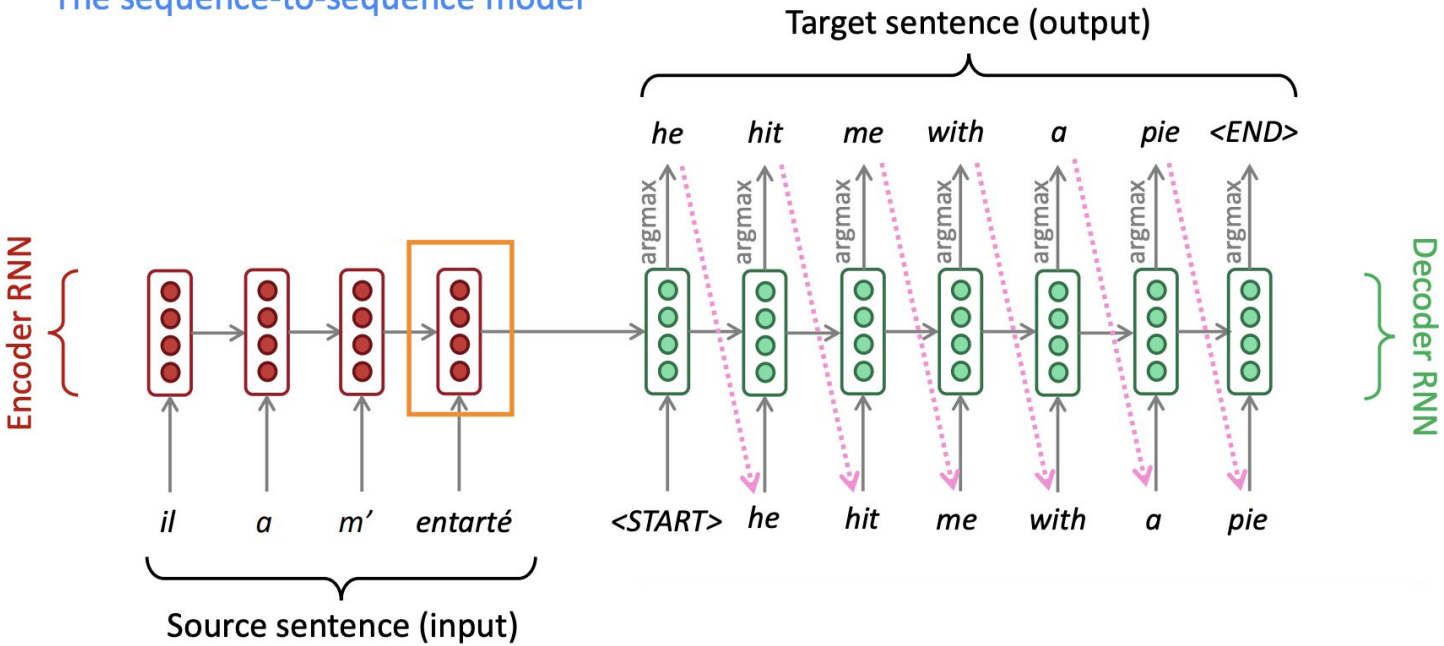
What is Neural Machine Translation?

- Neural Machine Translation (NMT) is a way to do Machine Translation with a *single neural network*

- The neural network architecture is called sequence-to-sequence (aka seq2seq) and it involves *two* RNNs.

Neural Machine Translation (NMT)

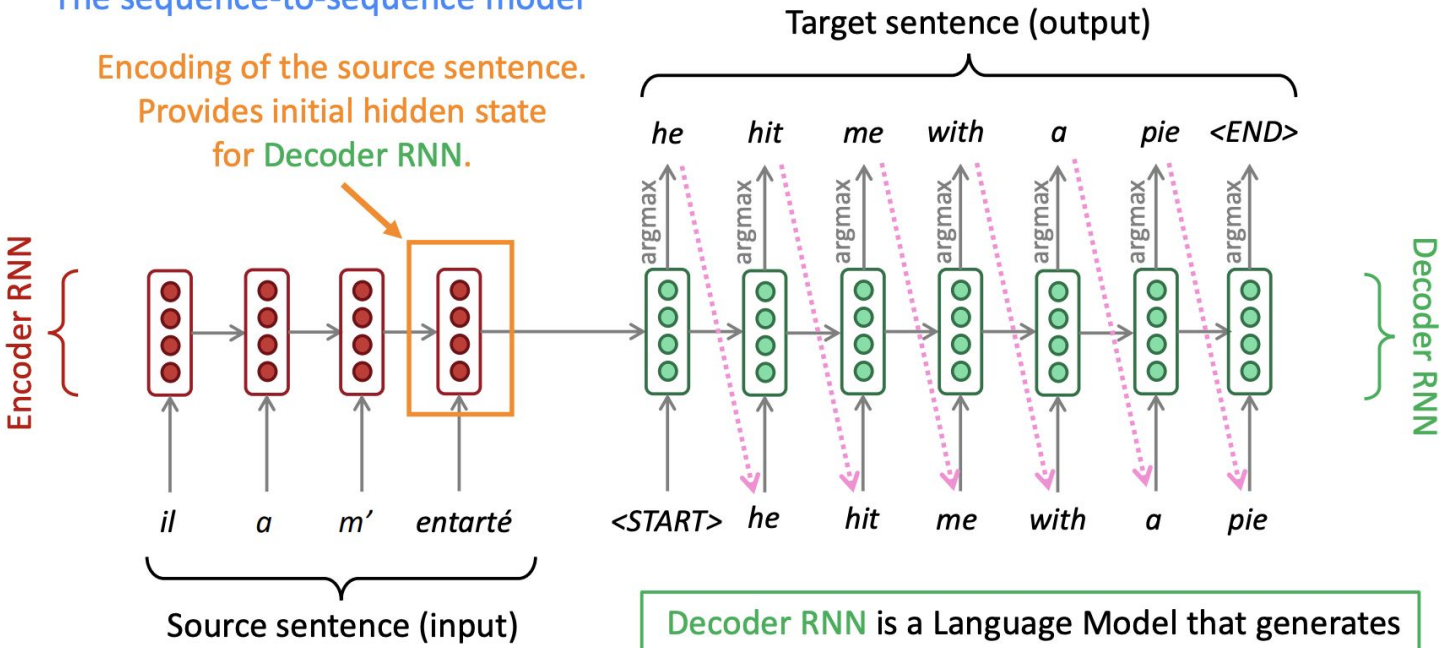
The sequence-to-sequence model



Encoder RNN produces an encoding of the source sentence.

Neural Machine Translation (NMT)

The sequence-to-sequence model

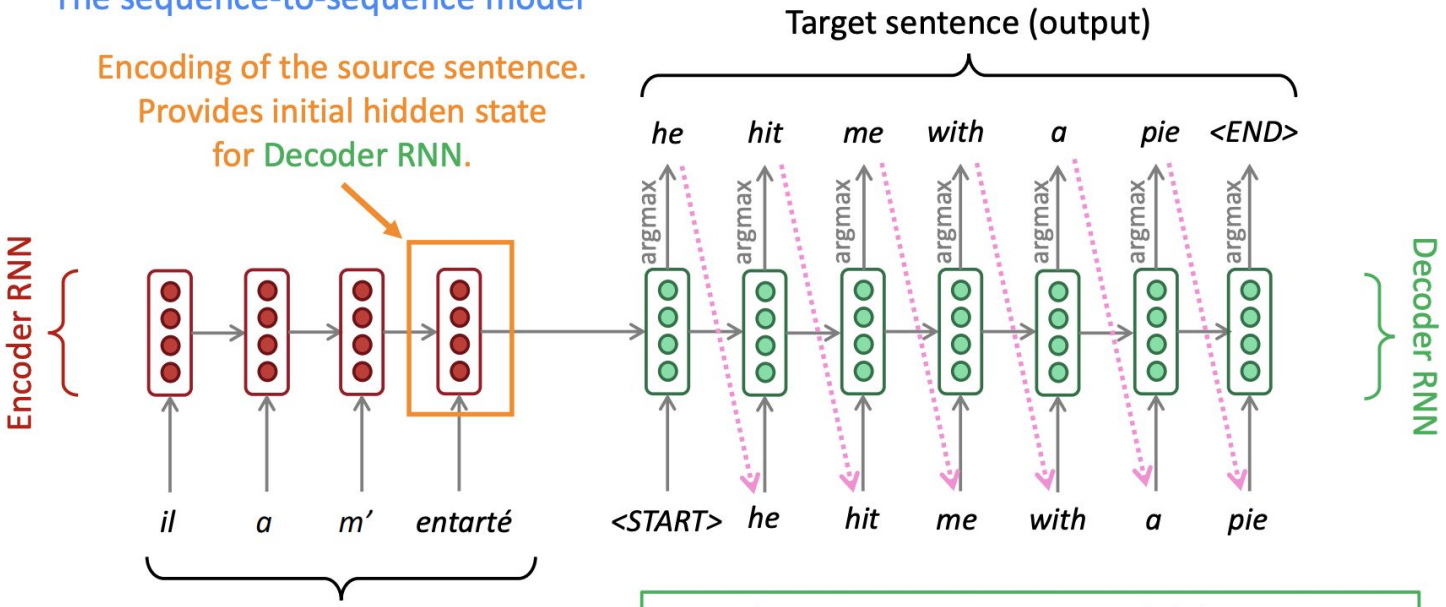


Decoder RNN is a Language Model that generates target sentence, *conditioned on encoding*.

Encoder RNN produces an encoding of the source sentence.

Neural Machine Translation (NMT)

The sequence-to-sequence model



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

Encoder RNN produces an encoding of the source sentence.

Note: This diagram shows **test time** behavior: decoder output is fed in ⋯→ as next step's input

Neural Machine Translation (NMT)

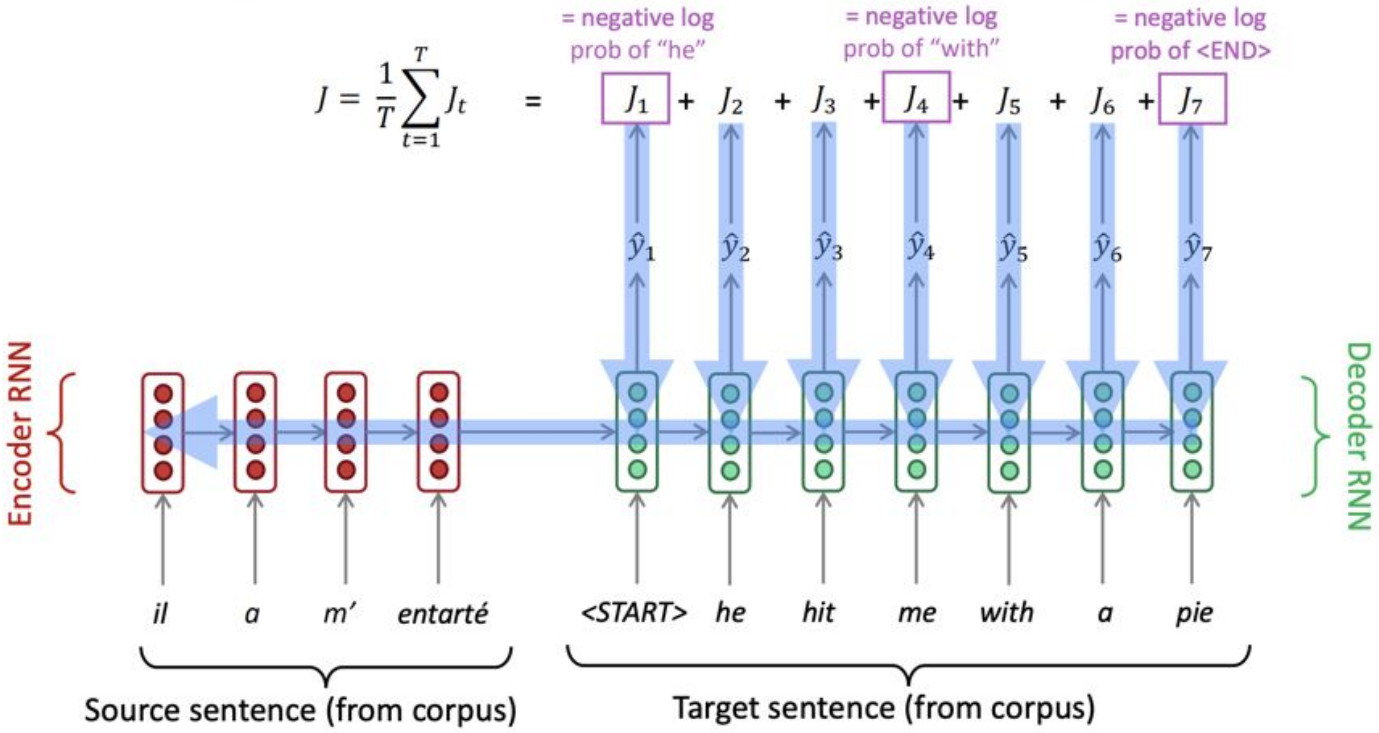
- The sequence-to-sequence model is an example of a **Conditional Language Model**.
  - **Language Model** because the decoder is predicting the next word of the target sentence *y*
  - **Conditional** because its predictions are *also* conditioned on the source sentence *x*

- NMT directly calculates $P(y|x)$:

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

Probability of next target word, given target words so far and source sentence *x*

Training a Neural Machine Translation system



Seq2seq is optimized as a **single system.**
Backpropagation operates "*end-to-end*".

Advantages of NMT

Compared to SMT, NMT has many advantages:

- Better performance
  - More fluent
  - Better use of context
  - Better use of phrase similarities
- A single neural network to be optimized end-to-end
  - No subcomponents to be individually optimized
- Requires much less human engineering effort
  - No feature engineering
  - Same method for all language pairs

Disadvantages of NMT?

Compared to SMT:

- NMT is less interpretable
  - Hard to debug
- NMT is difficult to control
  - For example, can't easily specify rules or guidelines for translation
  - Safety concerns!

How do we evaluate Machine Translation?

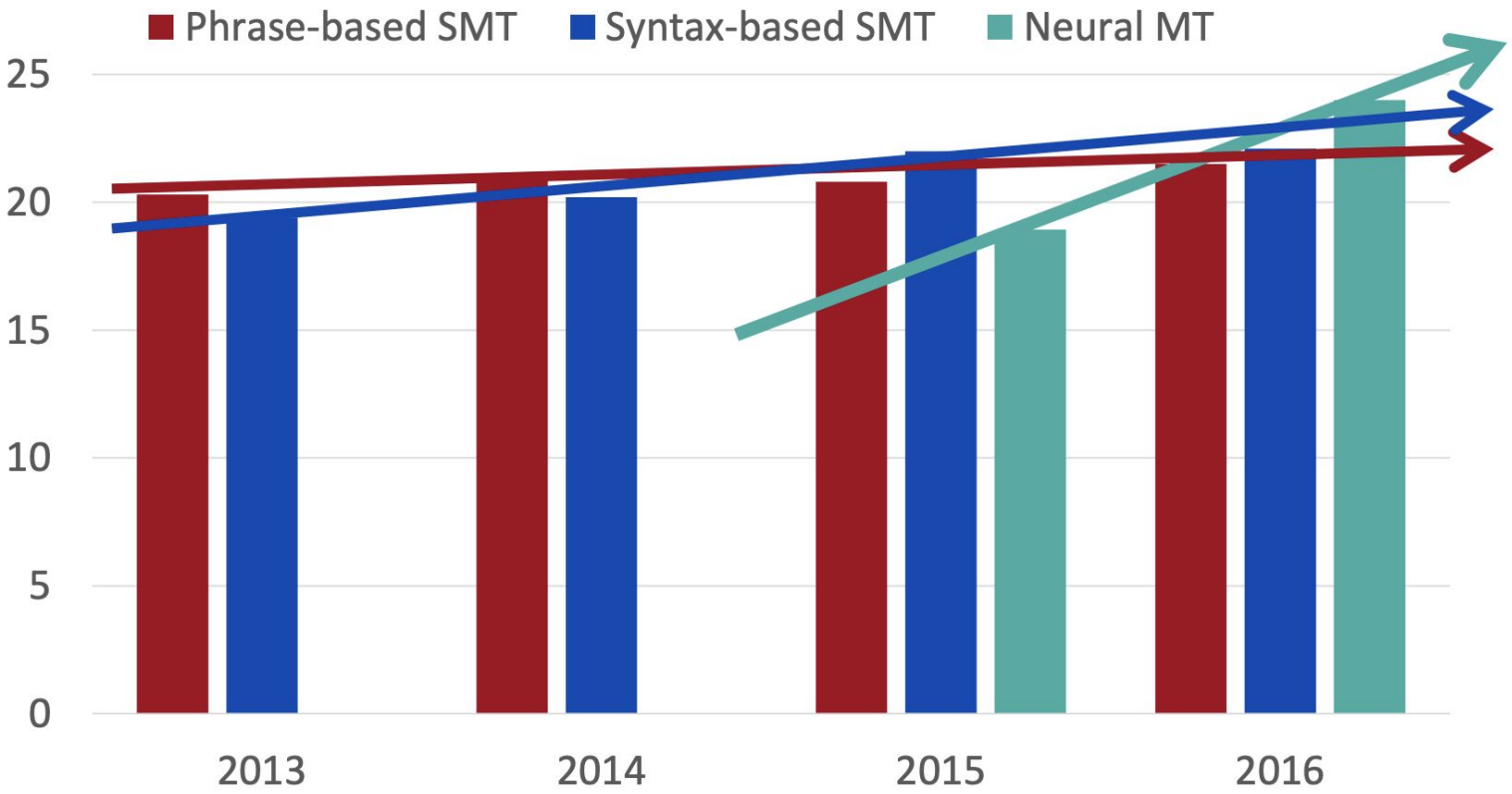
BLEU (**B**ilingual **E**valuation **U**nderstudy)

- BLEU compares the machine-written translation to one or several human-written translation(s), and computes a similarity score based on:
  - *n*-gram precision (usually for 1, 2, 3 and 4-grams)
  - Plus a penalty for too-short system translations

- BLEU is useful but imperfect
  - There are many valid ways to translate a sentence
  - So a good translation can get a poor BLEU score because it has low *n*-gram overlap with the human translation ☹

MT progress over time

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



NMT: the biggest success story of NLP Deep Learning

VietAI

Neural Machine Translation went from a fringe research activity in **2014** to the leading standard method in **2016**

- **2014**: First seq2seq paper published
- **2016**: Google Translate switches from SMT to NMT
- This is amazing!
  - **SMT** systems, built by hundreds of engineers over many years, outperformed by NMT systems trained by a handful of engineers in a few months

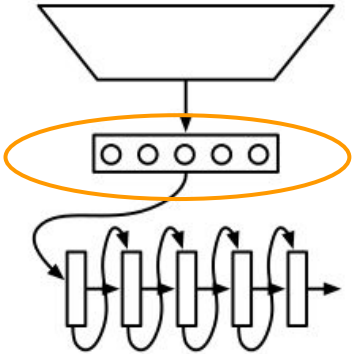
So is Machine Translation solved?

- **Nope!**
- Many difficulties remain:
  - Out-of-vocabulary words
  - Domain mismatch between train and test data
  - Maintaining context over longer text
  - Low-resource language pairs

2. Seq2Seq

NMT trước Seq2seq

x ***Kunst kann nicht gelehrt werden…***



$\mathbf{c}$

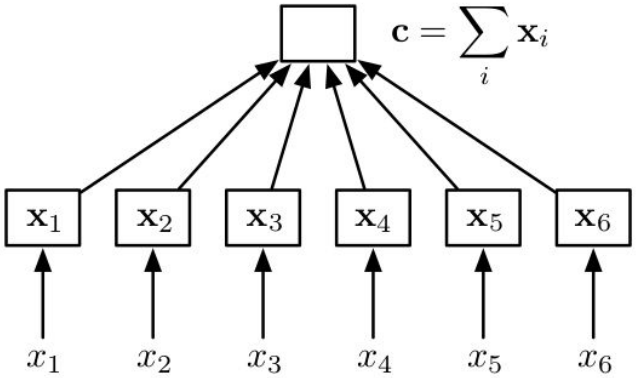
Encoder: Học 1 context vector

w ***Artistry can't be taught…***

NMT trước Seq2seq

How should we define $\mathbf{c} = \text{embed}(\boldsymbol{x})$?

The simplest model possible:



Encoder

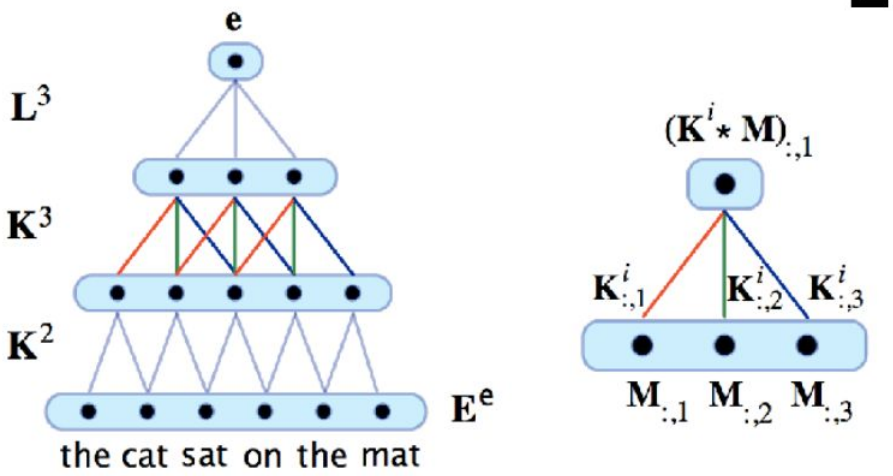
What do you think of this model?

NMT trước Seq2seq

How should we define $\mathbf{c} = \text{embed}(\boldsymbol{x})$?

Convolutional sentence model (CSM)

Encoder

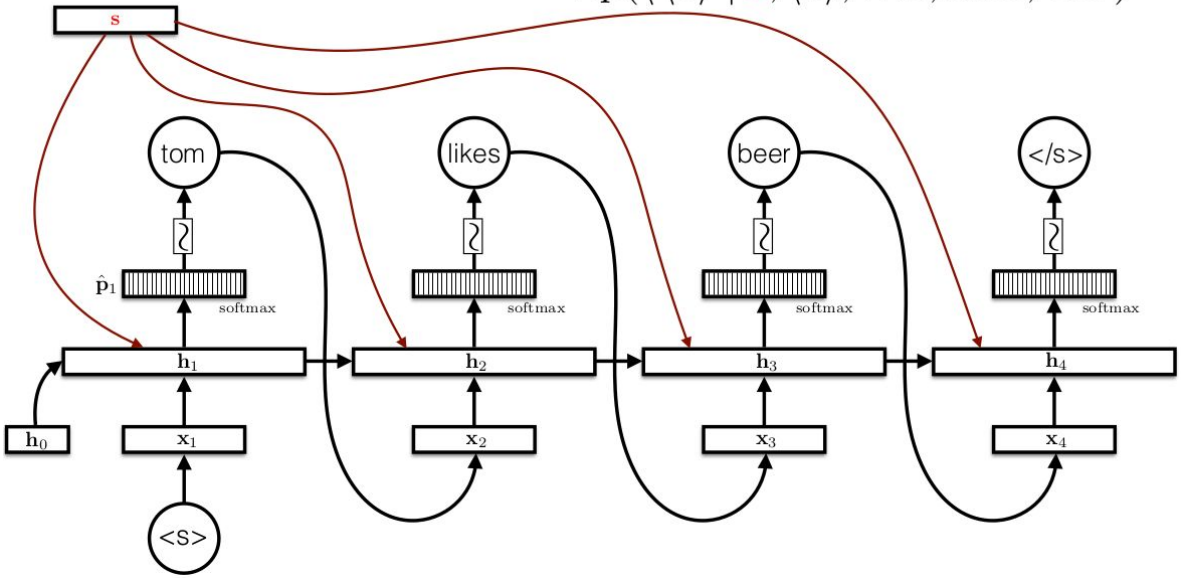


What do you think of this model?

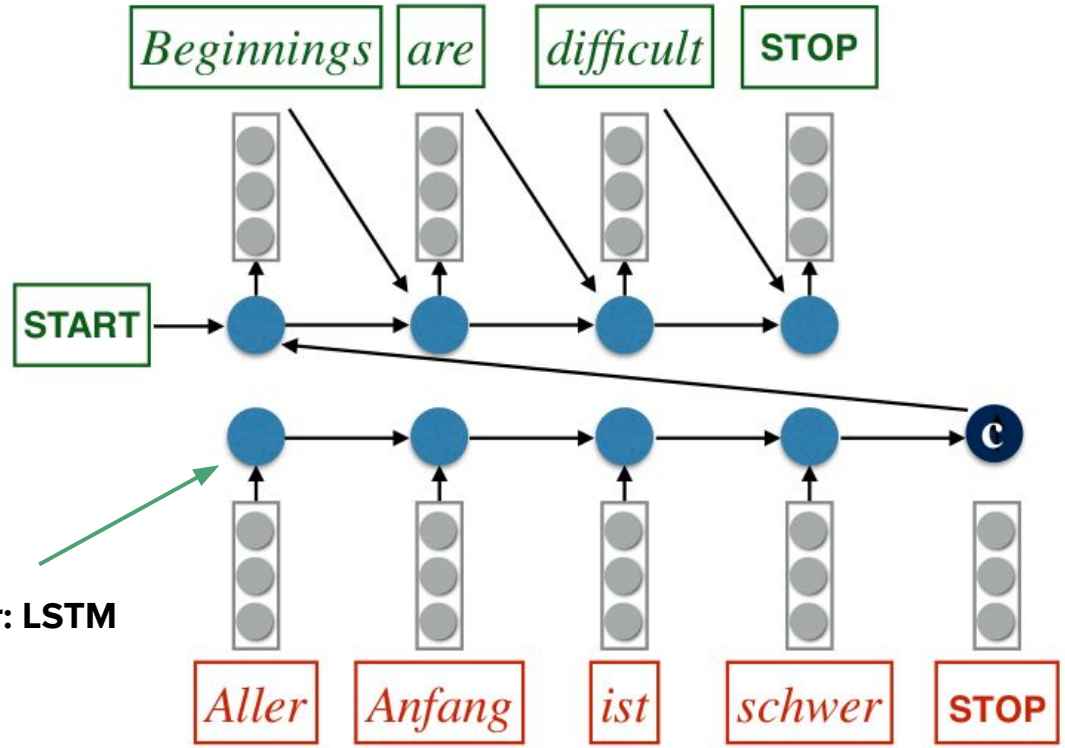
NMT trước Seq2seq

K&B 2013: RNN Decoder

$$p(tom \mid \mathbf{s}, \langle \mathbf{s} \rangle) \times p(likes \mid \mathbf{s}, \langle \mathbf{s} \rangle, tom) \times p(beer \mid \mathbf{s}, \langle \mathbf{s} \rangle, tom, likes) \times p(\langle \backslash \mathbf{s} \rangle \mid \mathbf{s}, \langle \mathbf{s} \rangle, tom, likes, beer)$$



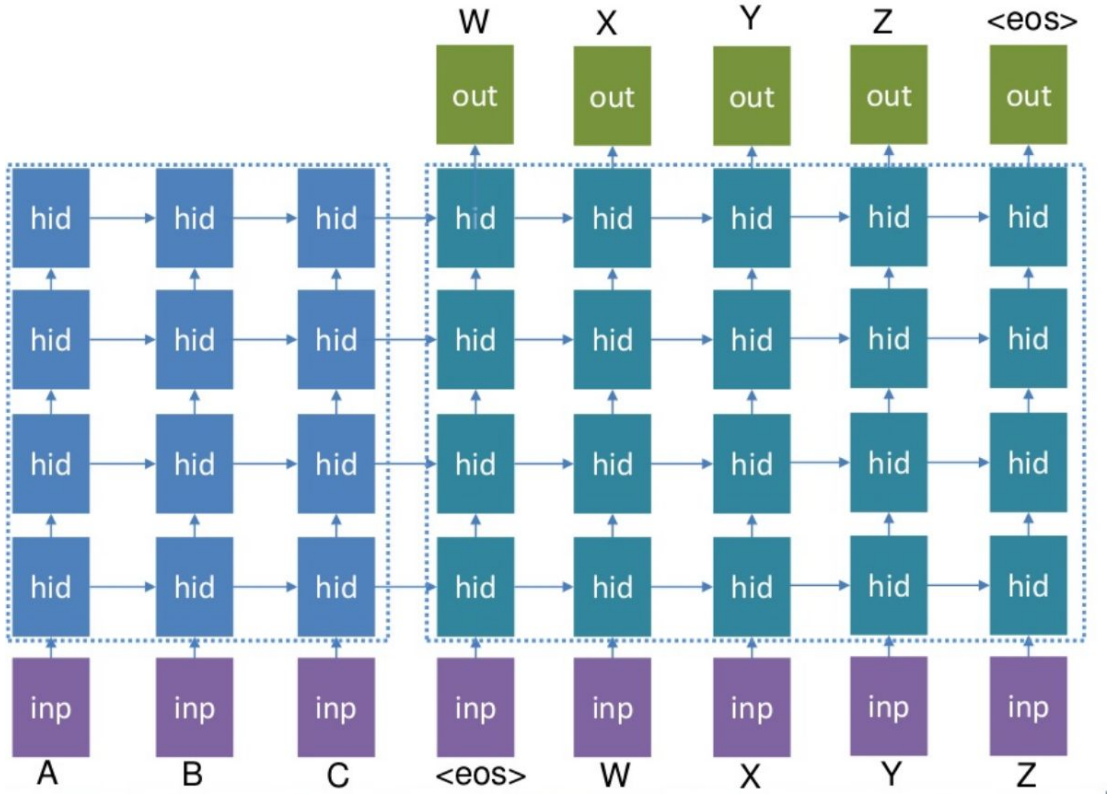
Seq2Seq



Original paper: LSTM

https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

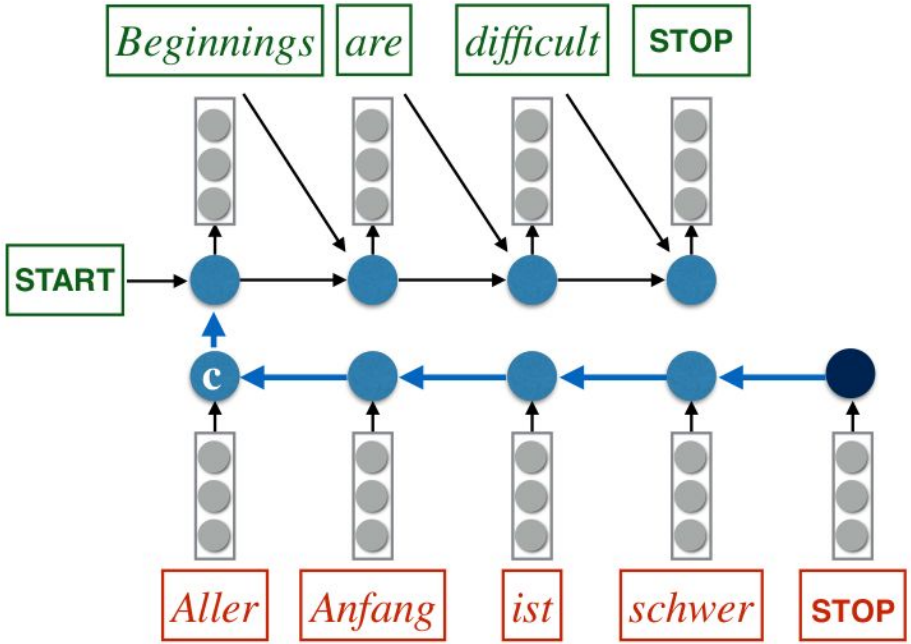
Seq2Seq: Go deeper



“We use minimum innovation to maximum results” - Ilya Sutskever

Seq2Seq: Trick

Read the input sequence “backwards”: **+4 BLEU**



Conditional Language model

$\boldsymbol{x}$ "input"	$\boldsymbol{w}$ "**text** output"
An author	A document written by that author
A topic label	An article about that topic
{SPAM, NOT_SPAM}	An email
A sentence in French	Its English translation
A sentence in English	Its French translation
A sentence in English	Its Chinese translation
An image	A text description of the image
A document	Its summary
A document	Its translation
Meterological measurements	A weather report
Acoustic signal	Transcription of speech
Conversational history + database	Dialogue system response
A question + a document	Its answer
A question + an image	Its answer

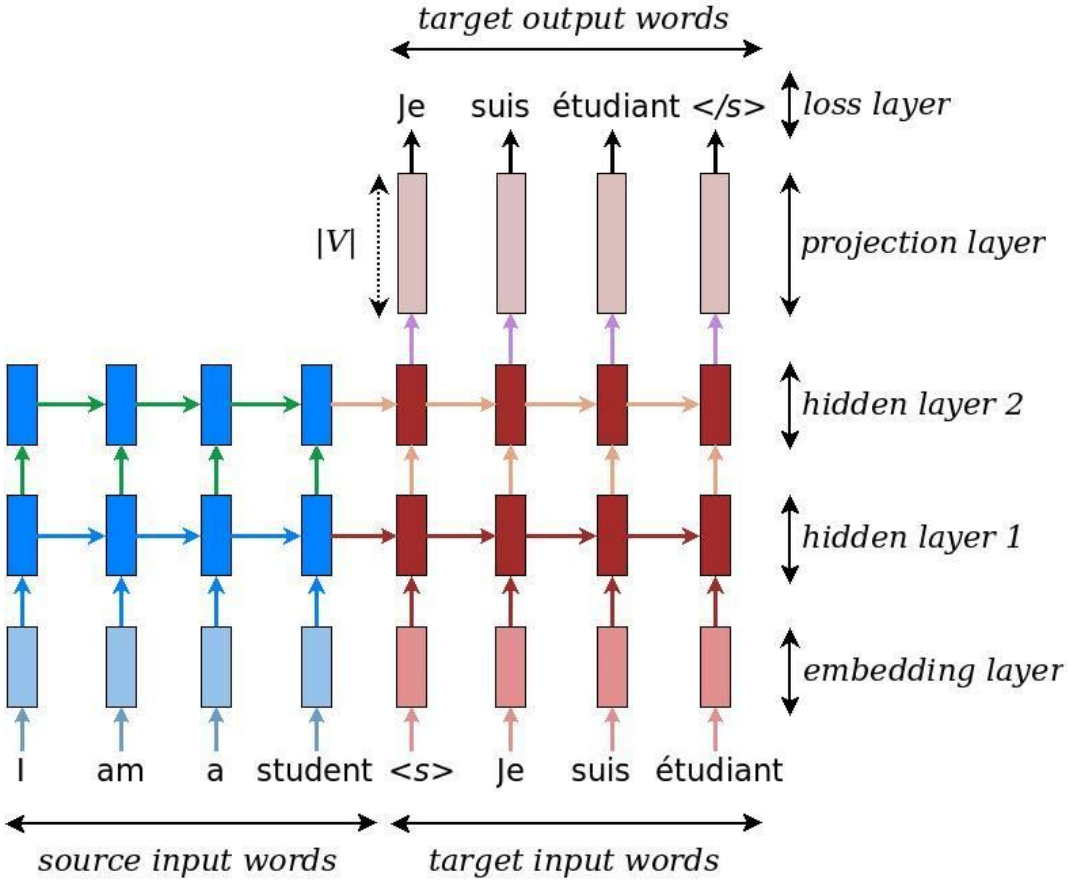
$$p(\boldsymbol{w} \mid \boldsymbol{x}) = \prod_{t=1}^{\ell} p(w_t \mid \boldsymbol{x}, w_1, w_2, \ldots, w_{t-1})$$

3. Một số ứng dụng thực tế của mô hình Seq2Seq

Dịch máy (Neural machine translation)

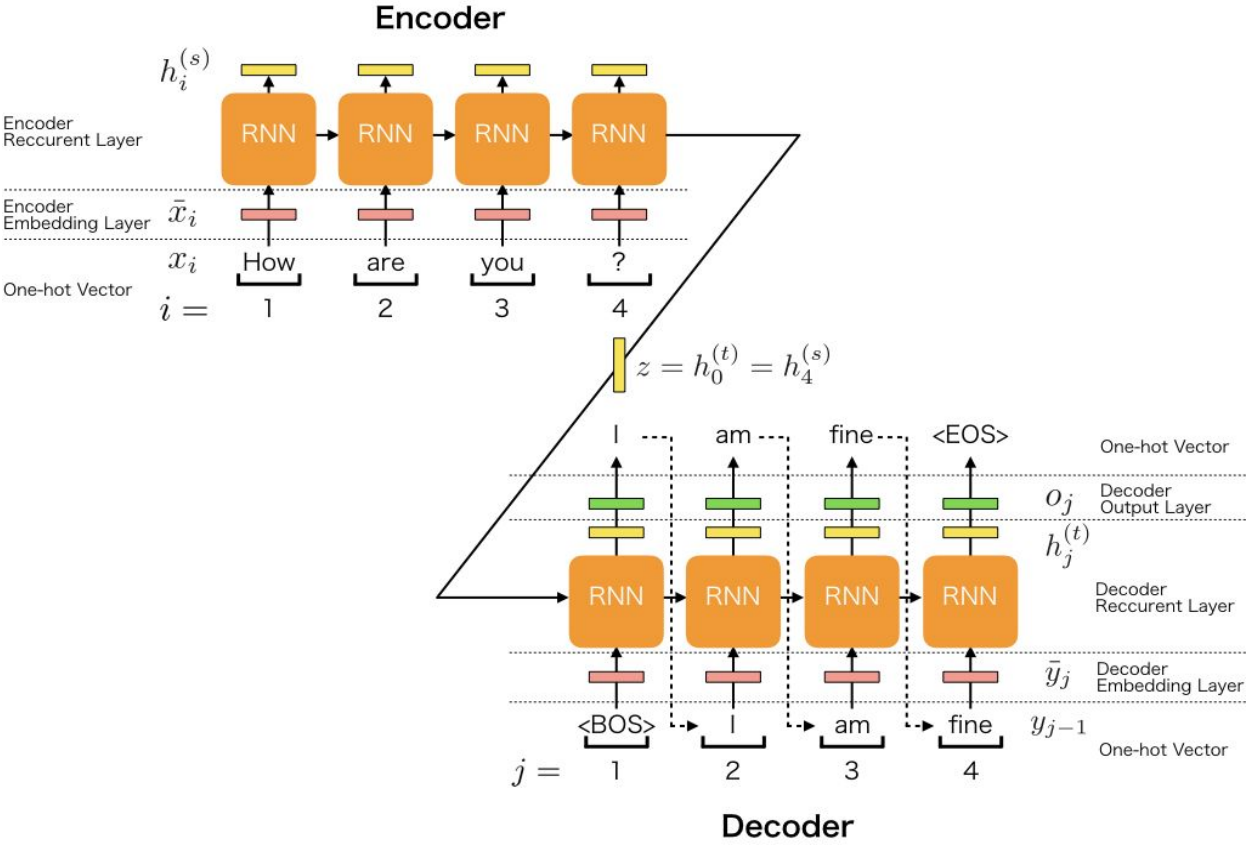
- **Input:** một câu (văn bản) trong ngôn ngữ nguồn (source).
- **Output:** một câu (văn bản) trong ngôn ngữ đích (destination).
- Bài toán kinh điển trong NLP (cả Understading lẫn Generation)
- Nhu cầu dịch văn bản qua lại giữa các ngôn ngữ là rất lớn
  - Google Translate dịch hơn 100 tỉ từ mỗi ngày
  - Facebook cũng đưa ra hệ thống dịch máy của riêng mình
  - eBay sử dụng dịch máy để phát triển thương mại xuyên quốc gia
  - Các phương pháp Machine Translation hiện nay đều dùng các ngữ liệu song ngữ (parallel corpus) để huấn luyện.

Neural machine translation



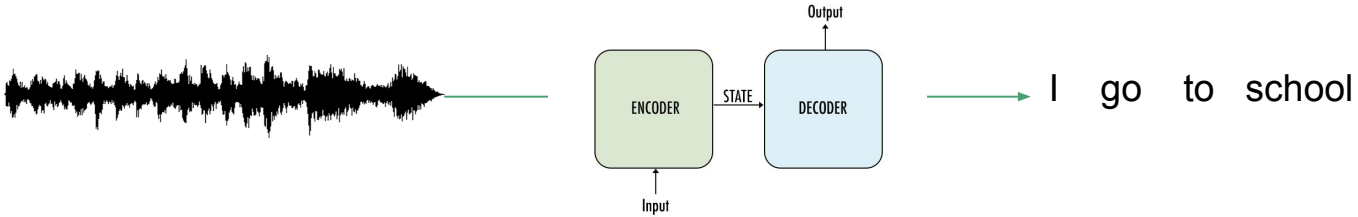
https://github.com/tensorflow/nmt

Chitchat conversational model



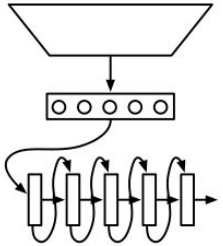
http://docs.chainer.org/en/stable/examples/seq2seq.html

Nhận dạng giọng nói

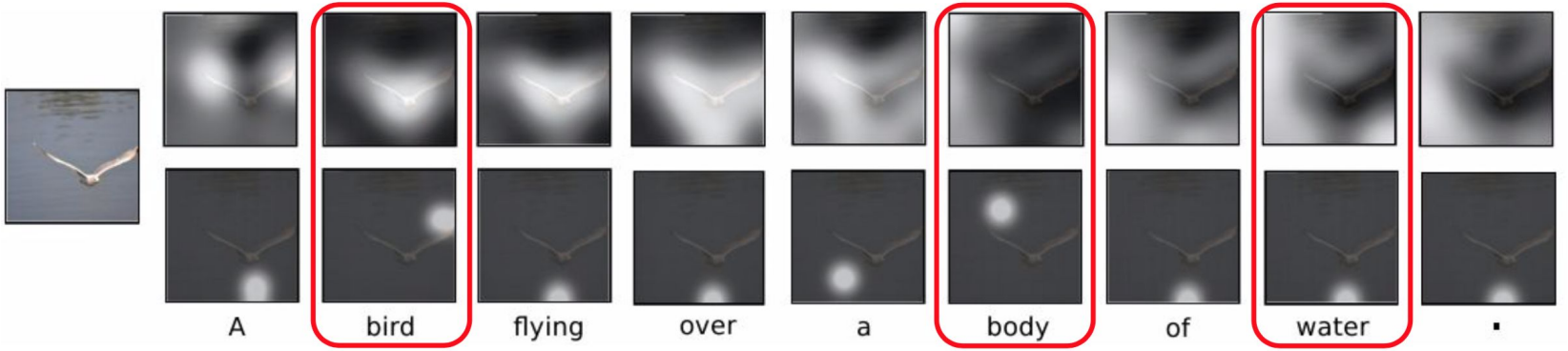


- Phương pháp truyền thống:
  - GMM-HMM: 30 years of feature engineering
  - DNN-GMM-HMM: Trained features
  - DNN-HMM: TDNN, LSTM, RNN
- End-to-end
  - Seq2seq

Image Captioning



A dog is playing on the beach.



Tại sao seq2seq lại trở nên rất phổ biến trong mọi bài toán

- Xây dựng được mô hình end2end, giảm thiểu độ phức tạp cho các task
- Đầu vào, đầu ra tuỳ ý (dễ dàng kết hợp nhiều loại dữ liệu từ text, hình ảnh, âm thanh,...)
- Có thể kết hợp nhiều task lại vào chung một mô hình (ví dụ dịch đa ngôn ngữ).

Q&A

Tài liệu tham khảo

1. CS224 http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture08-nmt.pdf
2. S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, Neural Computation 1997
3. Kyunghuyn Cho et al., Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, 2014
4. http://www.cs.toronto.edu/~graves/handwriting.html
5. http://www.cs.toronto.edu/~ilya/rnn.html