06 – MACHINE TRANSLATION

MACHINE LEARNING FOR NATURAL LANGUAGE PROCESSING, AIMS 2024

Lecture 06
Dr. Elvis Ndah

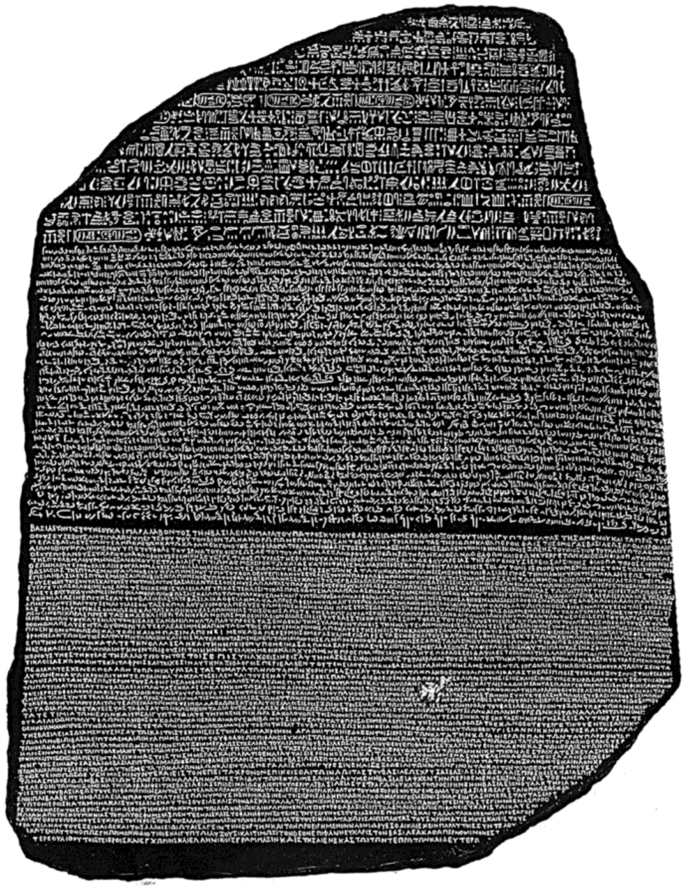
OVERVIEW

1. What is machine translation
2. Why is machine translation hard
3. Evaluation of machine translation
4. Encoder-decoder Architecture
5. Attention

THE ROSETTA STONE

First known historical evidence of translation

Instance of parallel text: Greek inscription allowed scholars to decipher the hieroglyphs



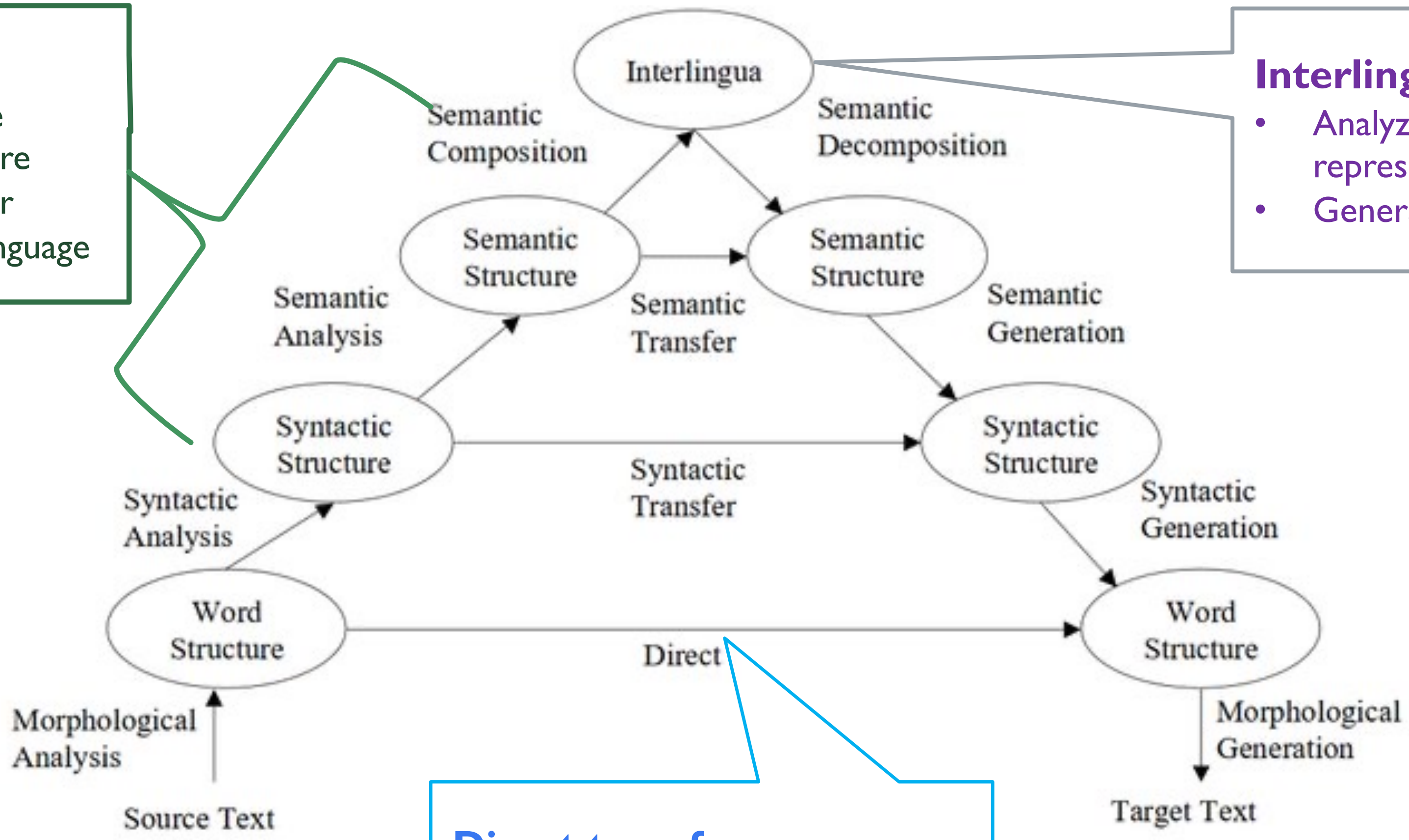
Hieroglyphic: used by priest in ancient Egypt

Demotic: used for daily purposes in Egypt

Ancient Greek: used by the administration

MACHINE TRANSLATION – THE VAUQUOIS TRIANGLE

Transfer-based:
- Parse source language
- Determine its structure
- Apply rules to transfer structure to target language



Interlingual:
- Analyze source language and represent as interlingual
- Generate target from interlingual

Direct transfer:
- word by word
- No language structure

MACHINE TRANSLATION (MT)

Machine Translation (MT)

- The task of translating a sentence *x* from one language (the source language) to another sentence *y* in another language (the target language).

X: L'homme est né libre, et partout il est dans les fers.

- Rousseau

Y: Man is born free, but everywhere he is in chains.

- Suppose we want to translate a text from *French* to *English*
- We need to find the *best English sentence* y, given a *French sentence* x

$$P(y|x), \forall y \in \Omega$$

MACHINE TRANSLATION (MT)

$$\underset{y}{\operatorname{argmax}}\, P(y|x) = \underset{y}{\operatorname{argmax}}\, \underbrace{P(x|y)}\underbrace{P(y)}$$

Bayes Rule

Translation Model

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

Language Model

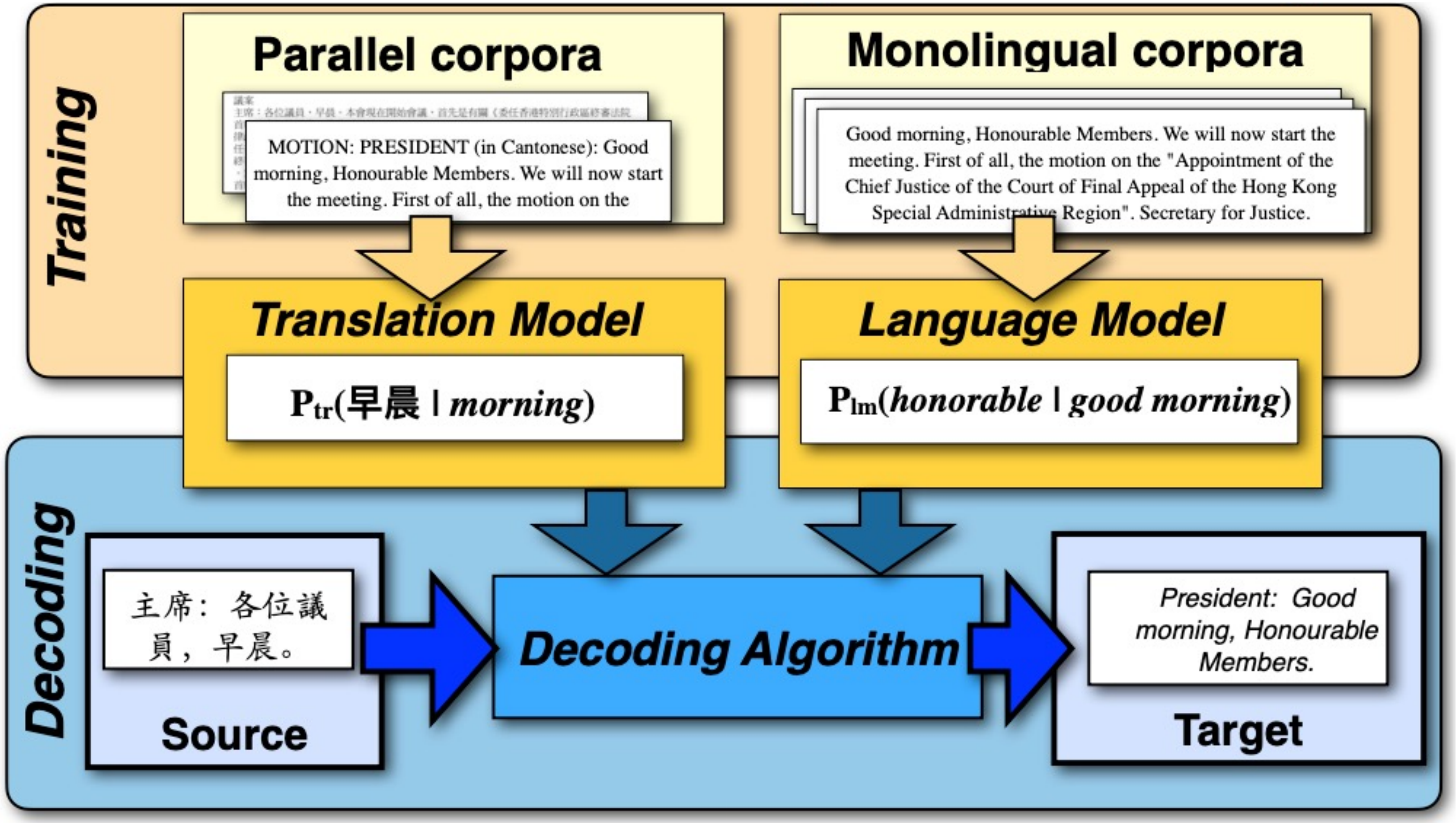
Models how to write good English (*fluency*). Learnt from monolingual data.

MACHINE TRANSLATION (MT)

$$\underbrace{\text{argmax}_y}_{\substack{\textbf{Question:}\\ \text{How to compute}\\ \text{this argmax?}}} \underbrace{P(x|y)}_{\text{Translation Model}} \underbrace{P(y)}_{\text{Language Model}}$$

- Enumerate every possible *y* and calculate the probability? Too expensive!
- **Solution :** *decoding*
  - Use a heuristic search algorithm to search for the best translation,
  - discarding hypotheses with very low-probability

STATISTICAL MACHINE TRANSLATION (SMT)



STATISTICAL MACHINE TRANSLATION (SMT)

How do we learn the translation model $P(x \mid y)$?

- large corpus of parallel text (target/source)
- Rewrite the translation model

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

where ***a*** is an alignment

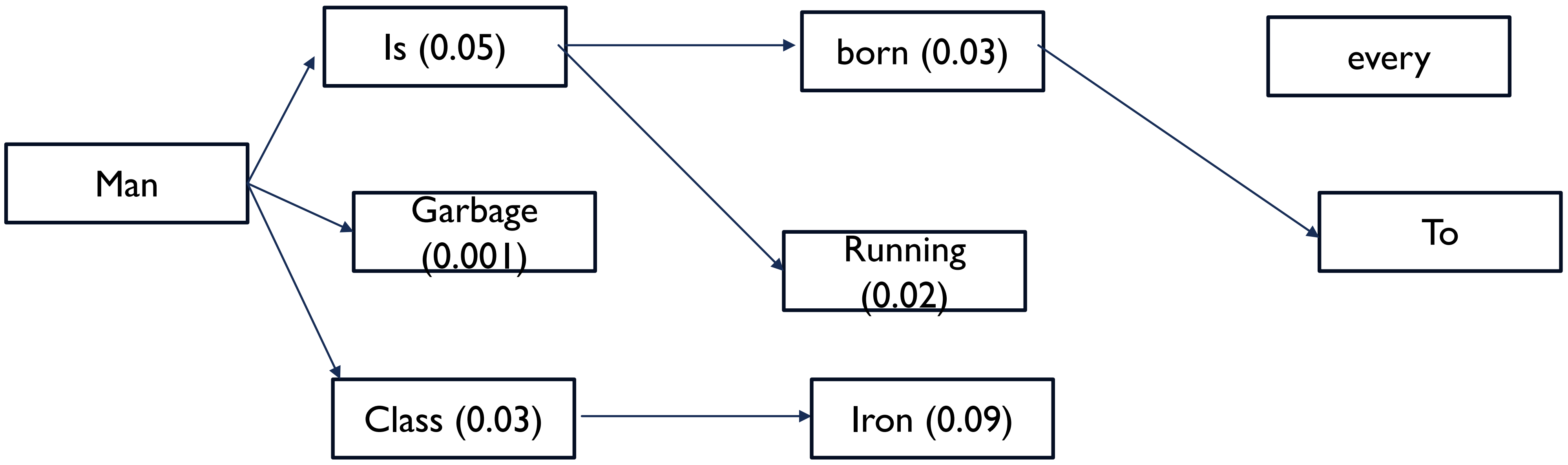
- an alignment is a correspondence between ***target (x)*** sentence and ***source (y)*** sentence
- The *alignment a* can be regarded as the decoder
- ***NOTE:*** *Obtaining an alignment (decoder) is not a trivial task*

STATISTICAL MACHINE TRANSLATION (SMT) – DECODING

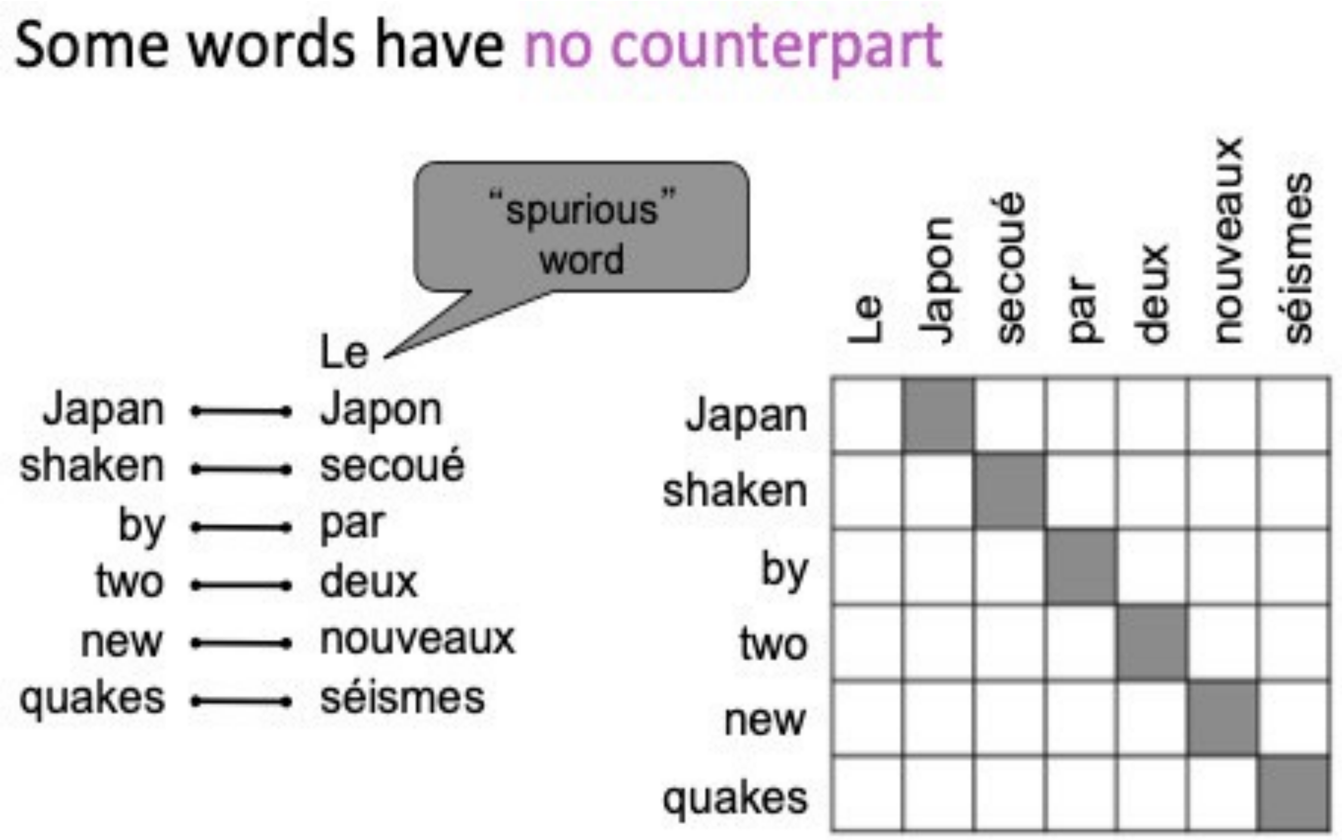
Find translation that maximizes $P(\, y \,|x)$

- **Exhaustive search decoding**
  - Try computing all possible sequences y (too expensive)
  - At each time step we are tracking V possible partial translations

- **Beam search decoding**
  - On each step of decoder keep track of the k most probable partial translation, with K the beam size
  - Beam search is not guaranteed to find optimal solution
  - More efficient than exhaustive search!


Man
Is (0.05)
Garbage (0.001)
Class (0.03)
born (0.03)
Running (0.02)
Iron (0.09)
every
To

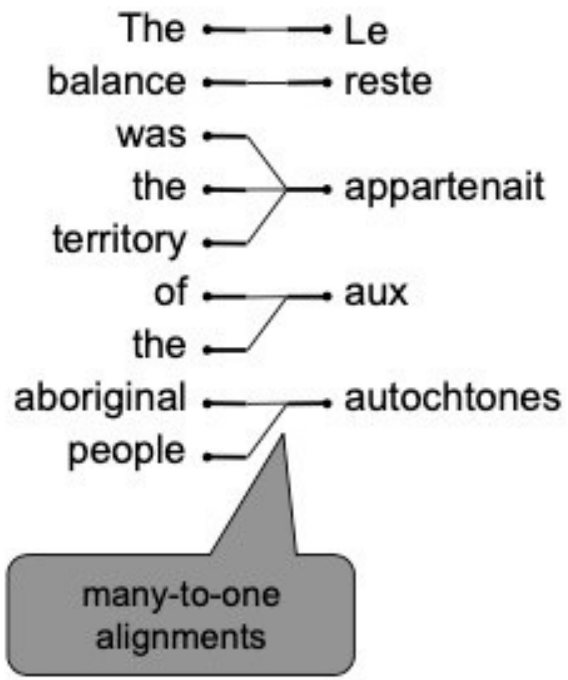
WHY IS MACHINE TRANSLATION HARD?

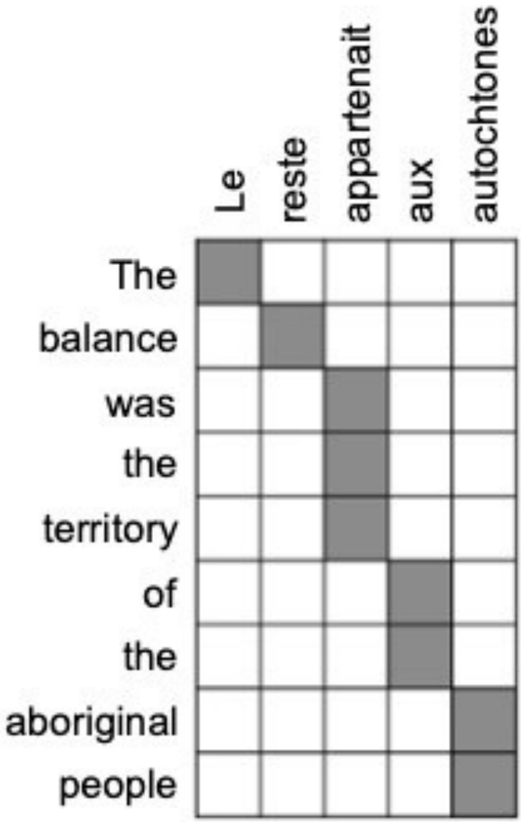


Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

WHY IS MACHINE TRANSLATION HARD?

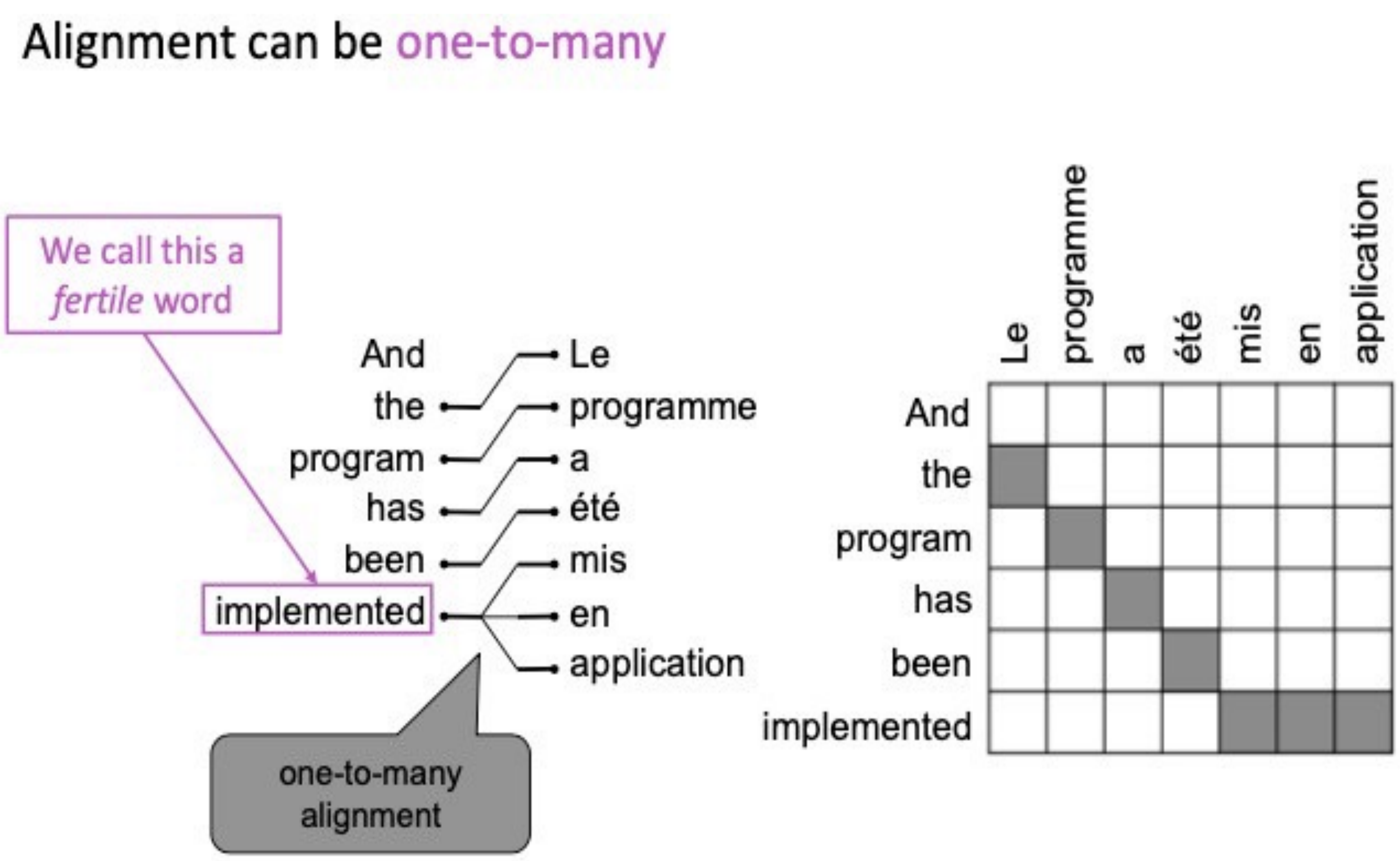
Alignment can be many-to-one





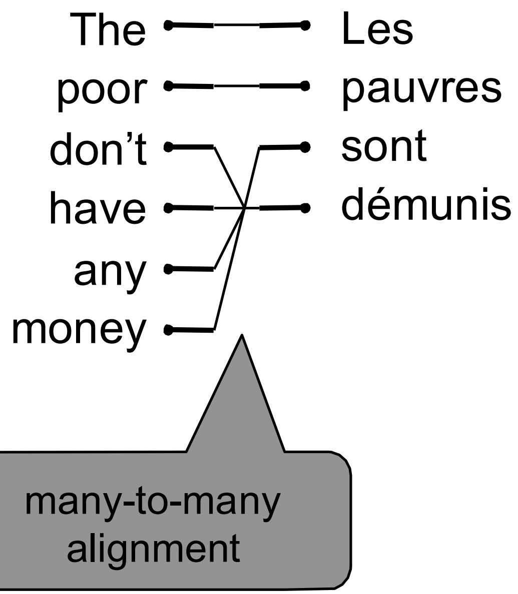
Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

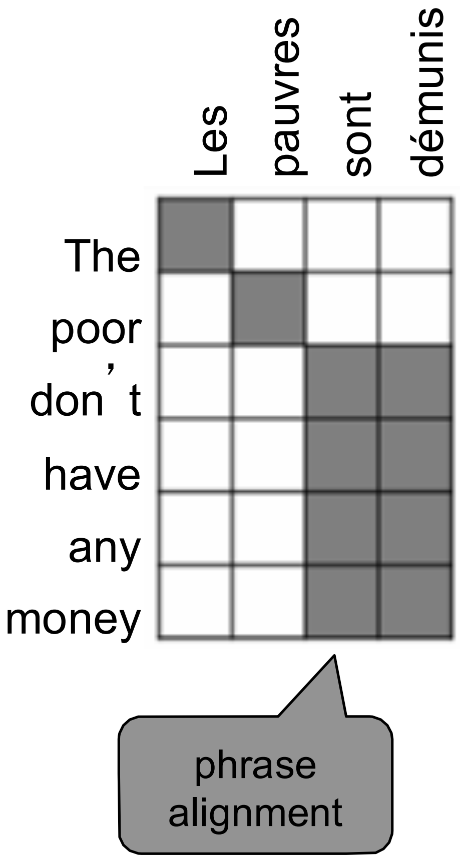
WHY IS MACHINE TRANSLATION HARD?



Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

WHY IS MACHINE TRANSLATION HARD?

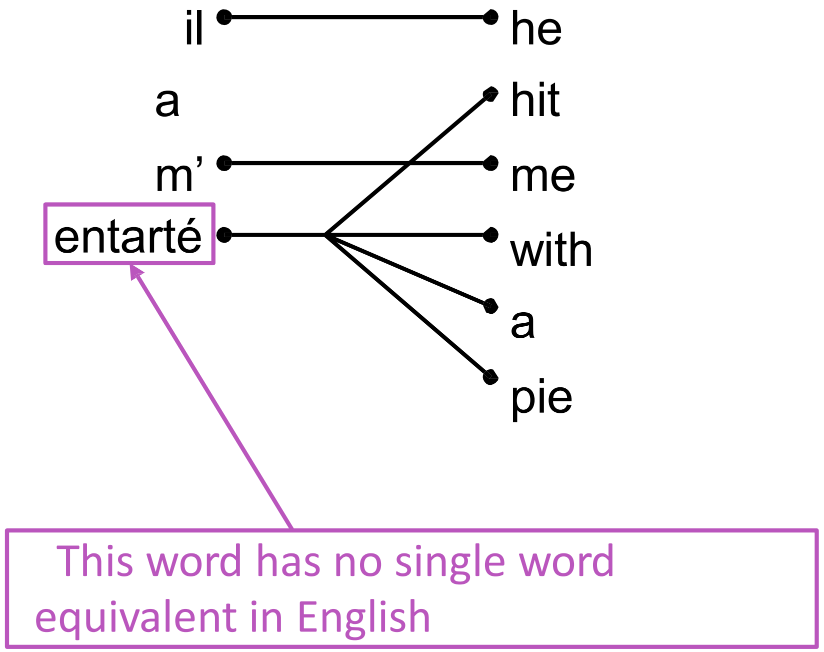


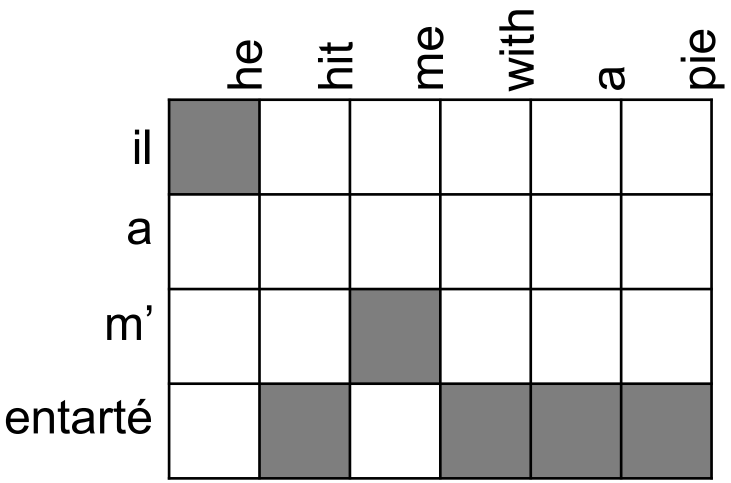


"The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

WHY IS MACHINE TRANSLATION HARD?

- Some words are very fertile! Can map multiple words in the same sentence





Examples from: “The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

MACHINE TRANSLATION – EVALUATION

What do we need to evaluate?

- Correctness of the translation
- Fluency of the translation, appropriateness
- We need appropriate evaluation metrics

Automatic evaluation:

- Inexpensive, can be done on a large scale, but may not capture what we want to evaluate.

Human evaluation:

- Expensive, and not easily reproducible or comparable across evaluations (different judges, different questions, ...)

AUTOMATIC EVALUATION – BLUE

BLUE: Bilingual Evaluation Understudy Score

- Evaluate candidate translations against several reference translations.
- The **BLUE** score is based on **n-gram** precisions'
  - How many n-grams in the candidate translation occur also in one of the reference translation.

C1: It is a guide to action which ensures that the military always obeys the commands of the party.
C2: It is to insure the troops forever hearing the activity guidebook that party direct

R1: It is a guide to action that ensures that the military will forever heed Party commands.
R2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
R3: It is the practical guide for the army always to heed the directions of the party.

AUTOMATIC EVALUATION – ISSUE WITH BLUE

What if some words are over-generated?

- **Example 1:**
  - Candidate: *the the the the the the the.*
  - Reference 1: *The cat is on the mat.*
  - Reference 2: *There is a cat on the mat.*
  - N-gram Precision: *7/7*

- **Solution:**
  - reference word should be exhausted after it is matched.

AUTOMATIC EVALUATION – ISSUE WITH BLUE

- **Example 2:**
  - Candidate: *the.*
  - Reference 1: *My mom likes the blue flowers.*
  - Reference 2: *My mother prefers the blue flowers.*
  - N-gram Precision: *1/1*

- **Solution:**
  - add a penalty if the candidate is too short.

AUTOMATIC EVALUATION – BLEU

Geometric Average

r = pick for each candidate in reference translation that is closest in length

$$\text{BLEU} = (p_1 \cdot p_2 \cdot p_3 \cdot p_4)^{\frac{1}{4}} \underbrace{\max(1,\ e^{1-\frac{r}{c}})}_{\text{Brevity Penalty}}$$

Clipped N-gram precisions for N=1, 2, 3, 4

c = length of the whole candidate translation corpus

- Ranges from 0.0 to 1.0, but usually shown multiplied by 100
- An increase of +1.0 BLEU is usually a conference paper
- MT systems usually score in the 10s to 30s
- Human translators usually score in the 70s and 80s

AUTOMATIC EVALUATION – BLUE ADVANTAGES

- Quick and inexpensive to calculate
- It is easy to understand
- It is language independent
- It correlates highly with human evaluation

HUMAN EVALUATION

We want to know whether the translation is **"good" and accurate** of the original.

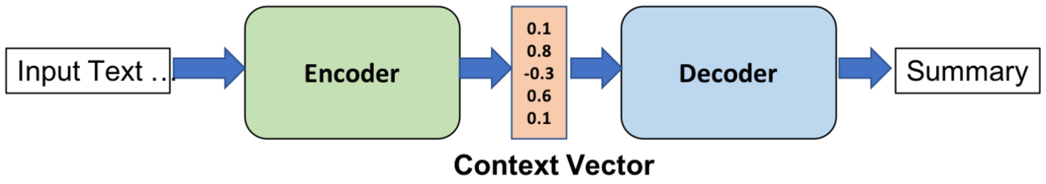
- Ask humans to judge the **fluency** and the **adequacy** of the translation
  - (e.g. on a scale of 1 to 5)

- Correlated with fluency is accuracy on **close task**:
  - Give raters the sentence with one word replaced by blank.
  - Ask raters to guess the missing word in the blank.

- Similar to adequacy is **informativeness**
  - Can you use the translation to perform some task
  - (e.g. answer multiple-choice questions about the text)

Encoder - decoder model

ENCODER – DECODER ARCHITECTURE



Encoder:

- at each time step take a single input of entire sequence.
- process the entire sequence and output a context vector

Context vector:

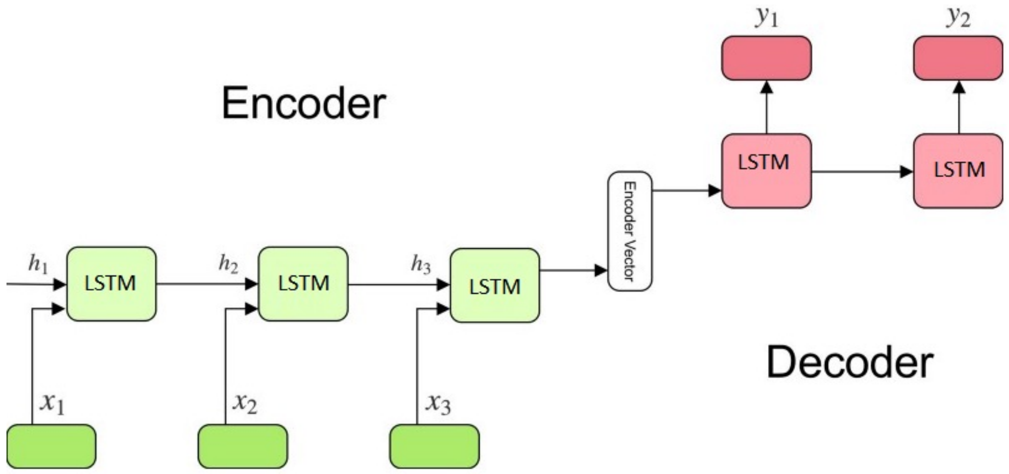
- conveys the essence of the input to the decoder.

Decoder:

- initialized from the final states of the Encoder (context vector).
- using initial states, decoder generates the output sequence.

ENCODER – DECODER ARCHITECTURE

- Consider a vocabulary *V*,
- an encoder-decoder is a function that maps a sequence $x = (x_1, \ldots, xn)$ onto another $y = (y_1, \ldots, ym)$



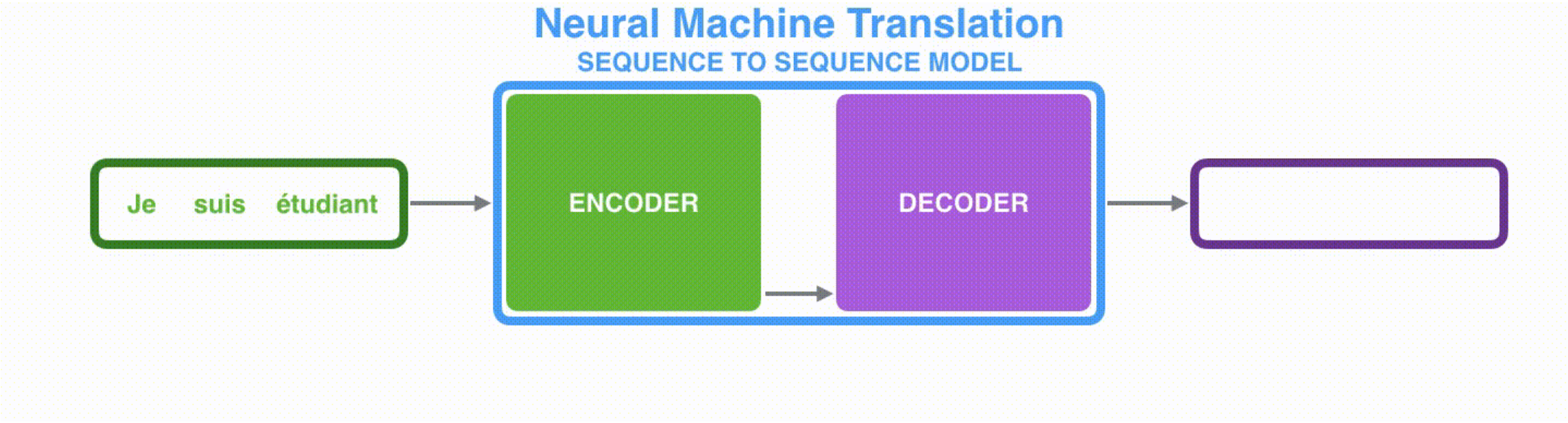
$$\mathbf{h}_t = g(\mathbf{h}_{t-1}, \mathbf{x}_t)$$
$$\mathbf{y}_t = f(\mathbf{h}_t)$$

- at time t the output y_t and hidden state h_t are computed as

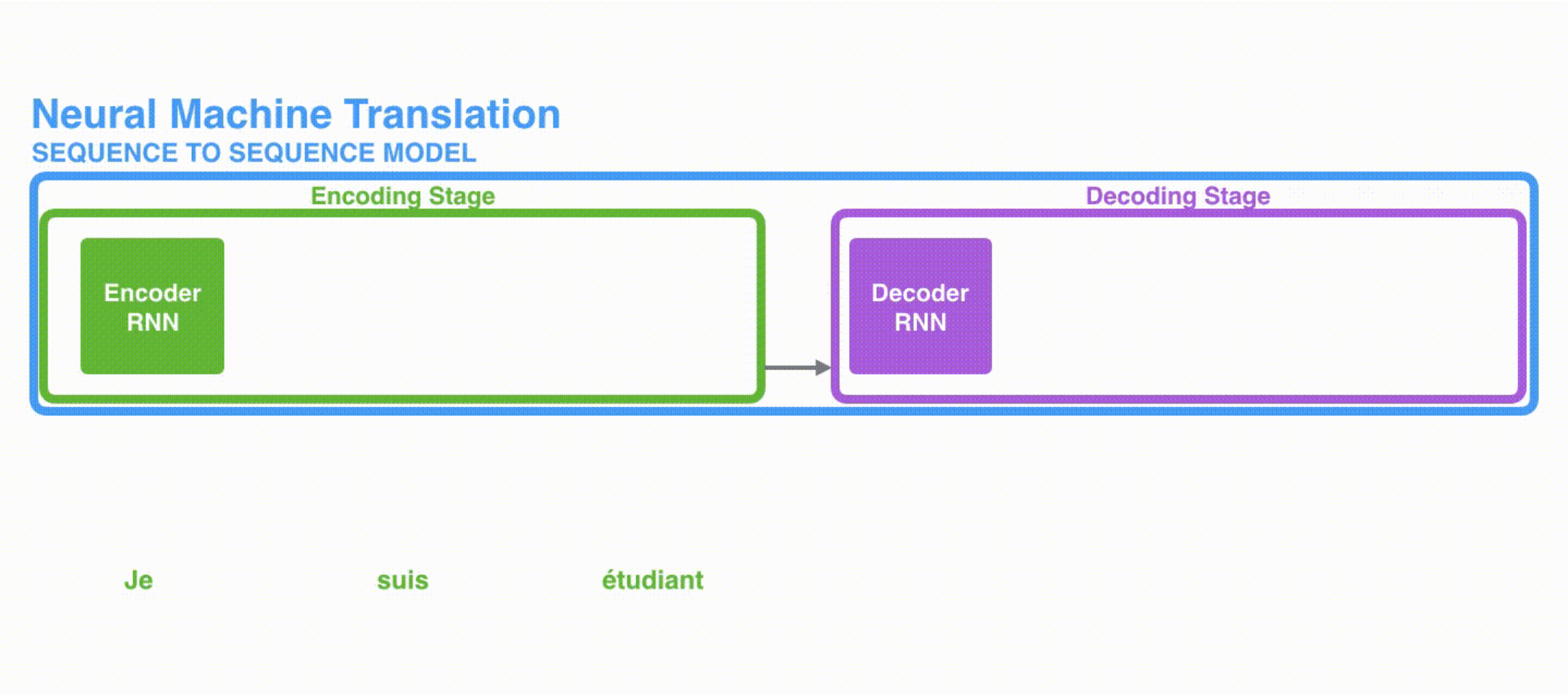
ENCODER – DECODER (SEQ2SEQ) MODEL

The model is composed of an encoder and a decoder.

- The encoder processes each item in the input sequence, it compiles the information it captures into a vector (called the ***context***).
- After processing the entire input sequence, the encoder sends the ***context*** over to the decoder, which begins producing the output sequence item by item.



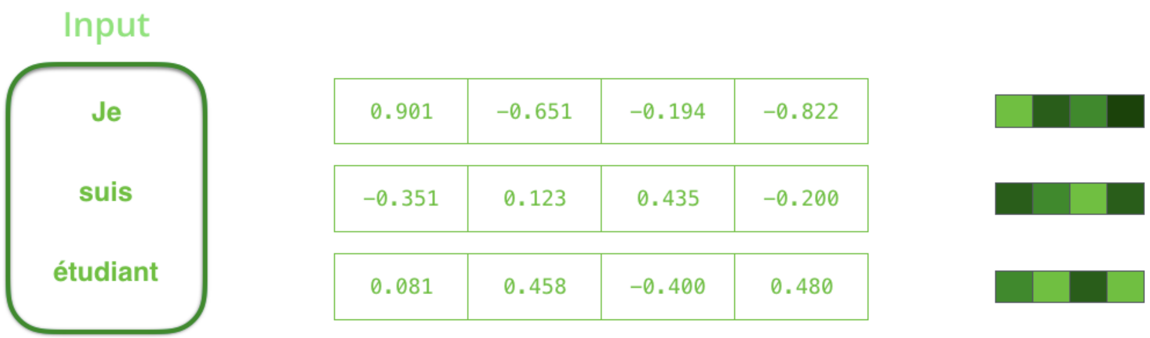
ENCODER – DECODER MACHINE TRANSLATION



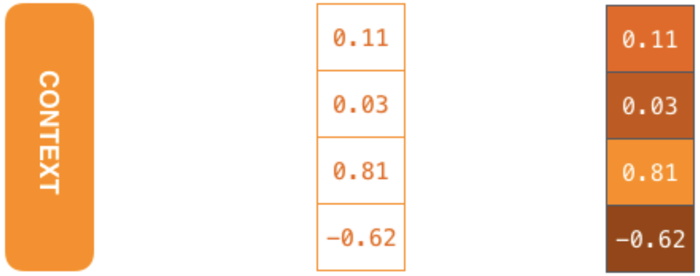
Credit: https://jalammar.github.io/

ENCODER – DECODER (SEQ2SEQ) MODEL

- **Input word:** Word embeddings



- **Context vector**
  - An array of real numbers with dimension the number of hidden units in the encoder (typical sizes are 256, 512 or 1024)



TRAINING AN ENCODER-DECODER MODEL

- training data typically consists of sets of sentences and their translations concatenated with a separator token.

- Encoder-decoder architectures are trained end-to-end, just as with the RNN language models. The network is given the source text and then starting with the separator token is trained autoregressively to predict the next word

INFERENCE FROM ENCODER-DECODER MODEL

Inference:

- During inference decoder uses its own estimated output y_t as the input for the next time step x_{t+1}.
- Thus, the decoder will tend to deviate more and more from the gold target sentence as it keeps generating more tokens

LIMITATIONS OF THE ENCODER – DECODER ARCHITECTURE

- **Weakness**
  the influence of the context vector (c) will wane as the output sequence is generate.

- **Solution:**
  make the context vector available at each step in the decoding process by adding it as a parameter to the computation of the current hidden state

DECODER – ENCODER ARCHITECTURE

- The context vector turned out to be a bottleneck for these types of models.
- Its challenging for the models to deal with long sentences.

- **Solution:**
  - Attention
    - Bahdanau et al., 2014 introduced
    - Luong et al., 2015. refined
  - Attention allows the model to focus on the relevant parts of the input sequence as needed.
    - highly improved the quality of machine translation systems.

Attention

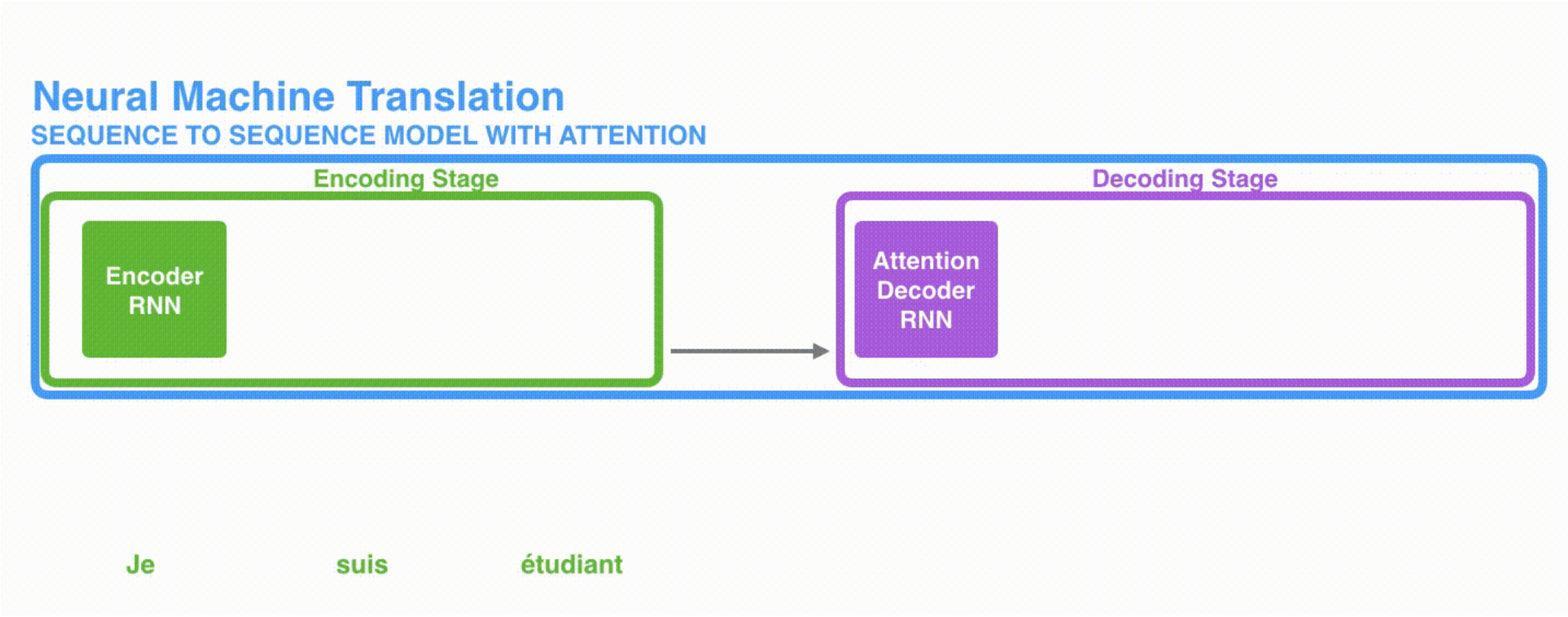
ATTENTION

“One important property of human perception is that one does not tend to process a whole scene in its entirety at once. Instead humans focus attention selectively on parts of the visual space to acquire information when and where it is needed, and combine information from different fixation over time to build up an internal representation of the scene, guiding future eye movements and decision making.”

- *Recurrent Models of visual Attention*

HOW ATTENTION DIFFER FROM CLASSIC SEQ2SEQ MODEL

1. the encoder passes *all* the hidden states (context) to the decoder. Instead of passing the last hidden state of the encoding stage, the encoder passes *all* the hidden states to the decoder



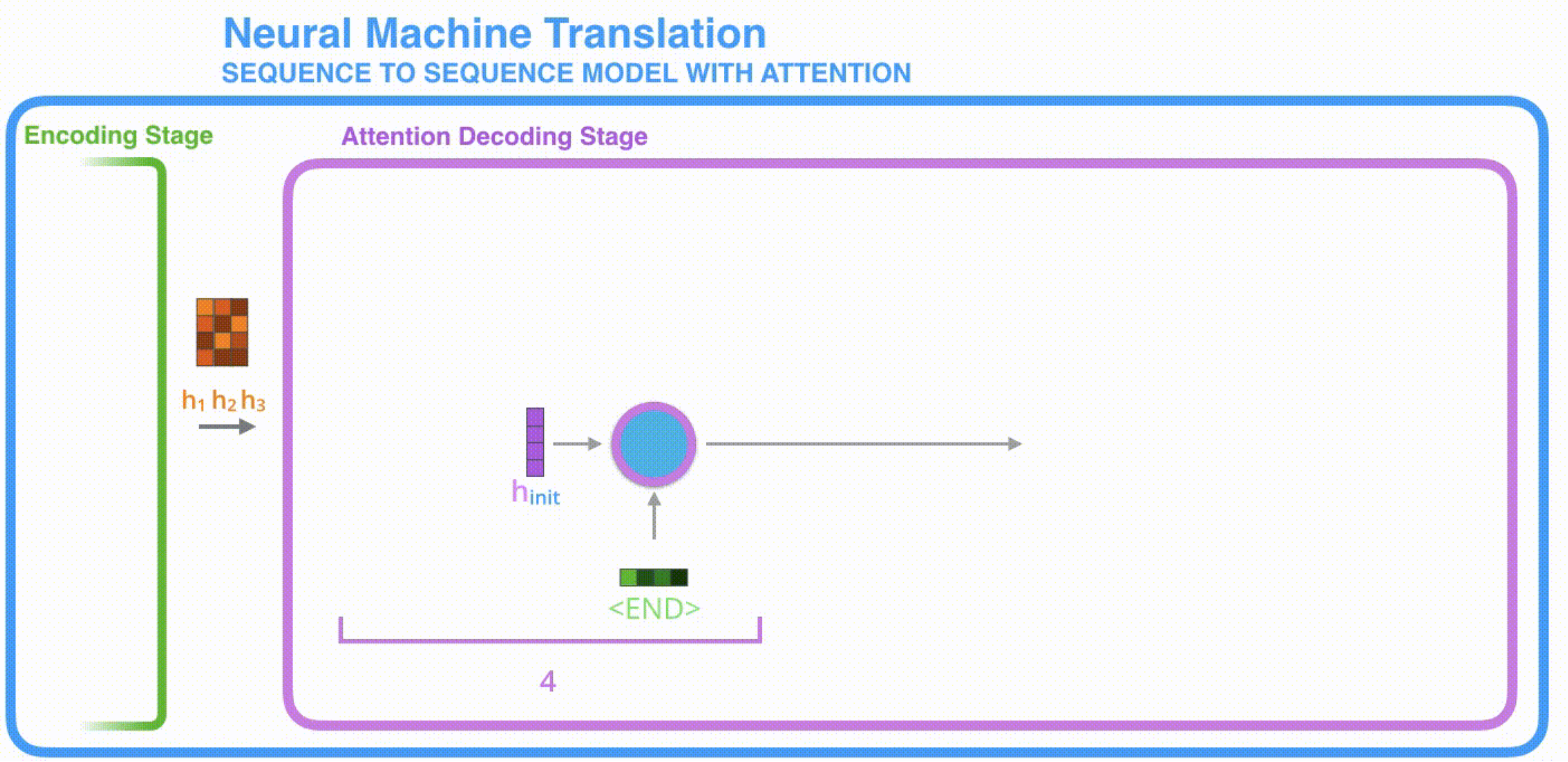
HOW ATTENTION DIFFER FROM CLASSIC SEQ2SEQ MODEL

2. To focus on the relevant parts of the input decoder does the following
   i. Evaluate each encoder hidden states – each encoder hidden states is most associated with a certain word in the input sentence.
   ii. Assign a score to each hidden states (more later)
   iii. Multiply each hidden states by its softmaxed score, thus amplifying hidden states with high scores, and drowning out hidden states with low scores.

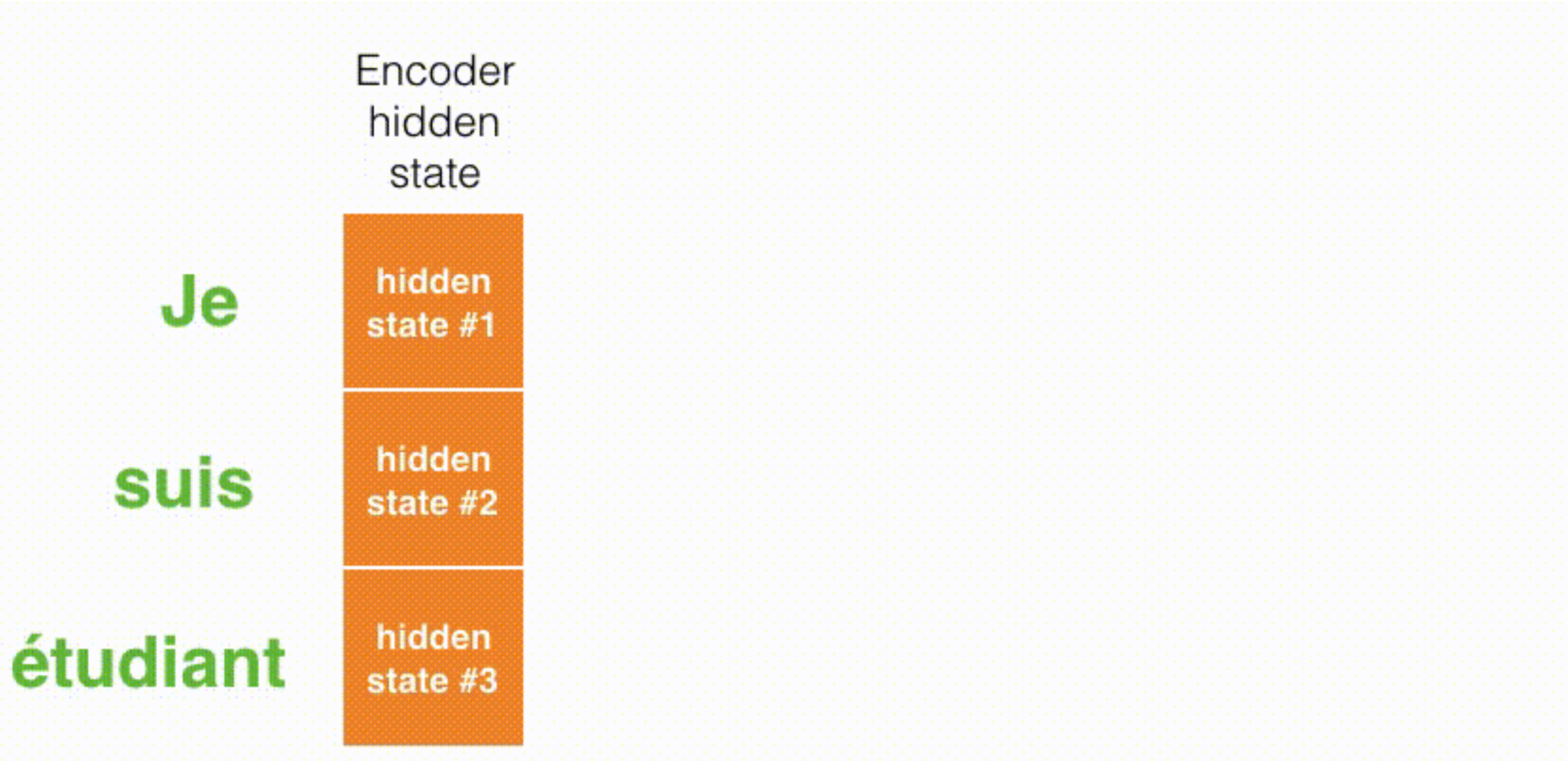
Attention: Encoding process

Attention

Scoring is done at each timestep on the decoder side



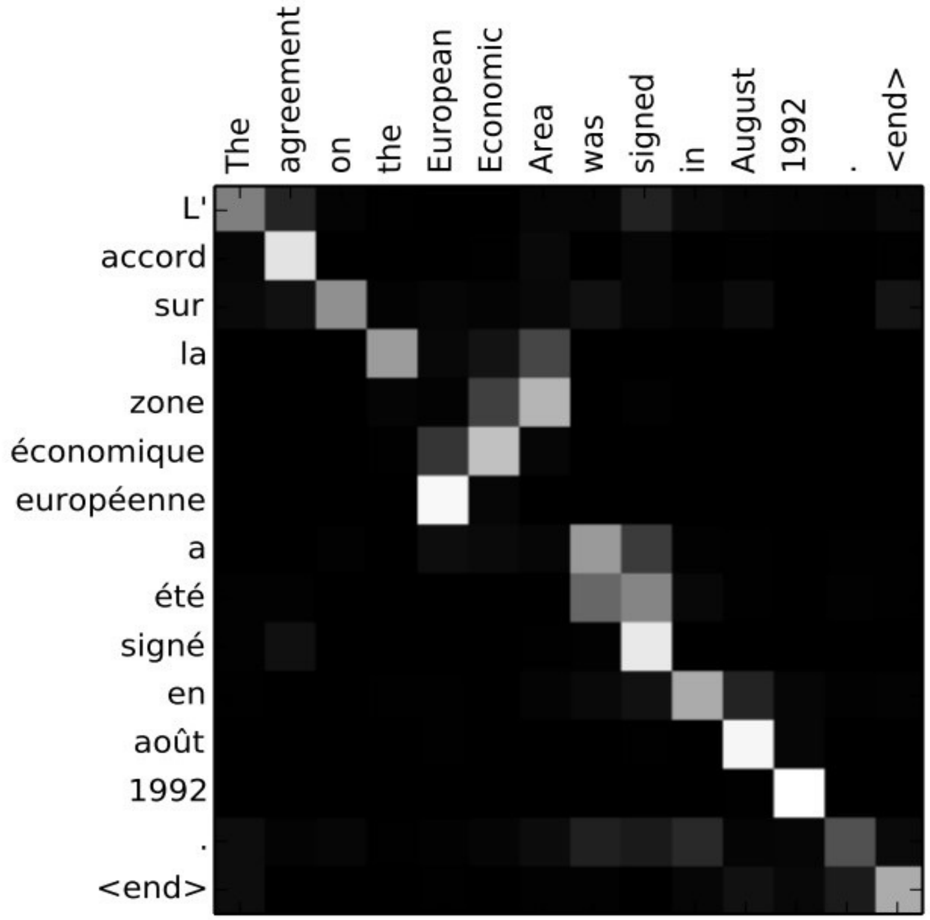
VISUALIZE ATTENTION



ATTENTION

Example of precision of the attention mechanism.

- The model learns how to align words in the language pair
- You can see how the model paid attention correctly
    - *European Economic Area*
    - *zone européenne économique*
    - Every other word in the sentence is in similar order.



REFERENCES

- Isutskeverb et al. (NIPS 2014) . Sequence to Sequence Learning with Neural Networks.
- https://www.scaler.com/topics/deep-learning/sequence-to-sequence-model/
- Vaswani et al. (NIPS 2017). Attention Is All You Need.