Lecture 17: Machine Translation

Zhizheng Wu

Agenda

- Recap
- Introduction
- Why machine translation is difficult
- Statistical MT
- Neural MT

Probability of next word

$$P(\text{best} \mid \text{Students from my class are the}) = \frac{C(\text{Students from my class are the best})}{C(\text{Students from my class are the})}$$

- C(Students from my class are the best) is count of the phrase "*Students from my class are the best*"

Generalizing bigram to n-gram

- From bigram to n-gram

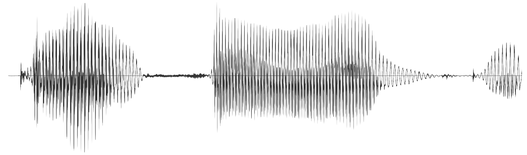
$$P(w_n \mid w_{1:n-1}) \approx P(w_n \mid w_{n-N+1:n-1})$$

- N = 2: bigram
- N = 3: trigram
- N = 4: 4-gram
- N = 5: 5-gram

TTS

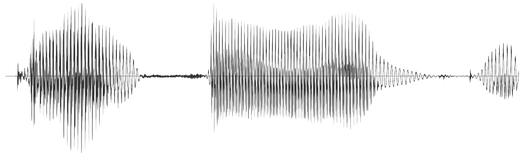
text → ***Text-to-Speech*** → *waveform*

`Author of the…`

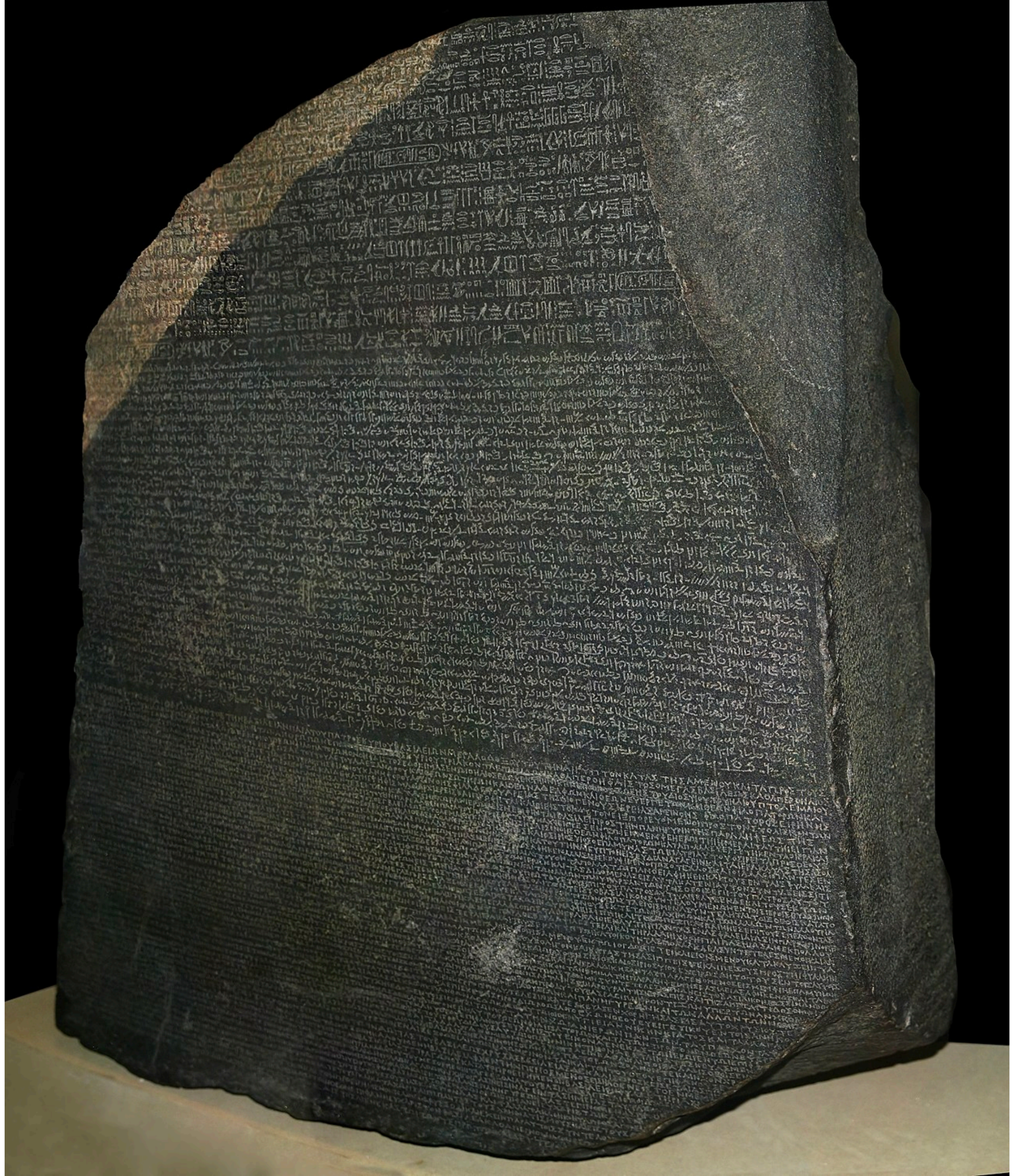


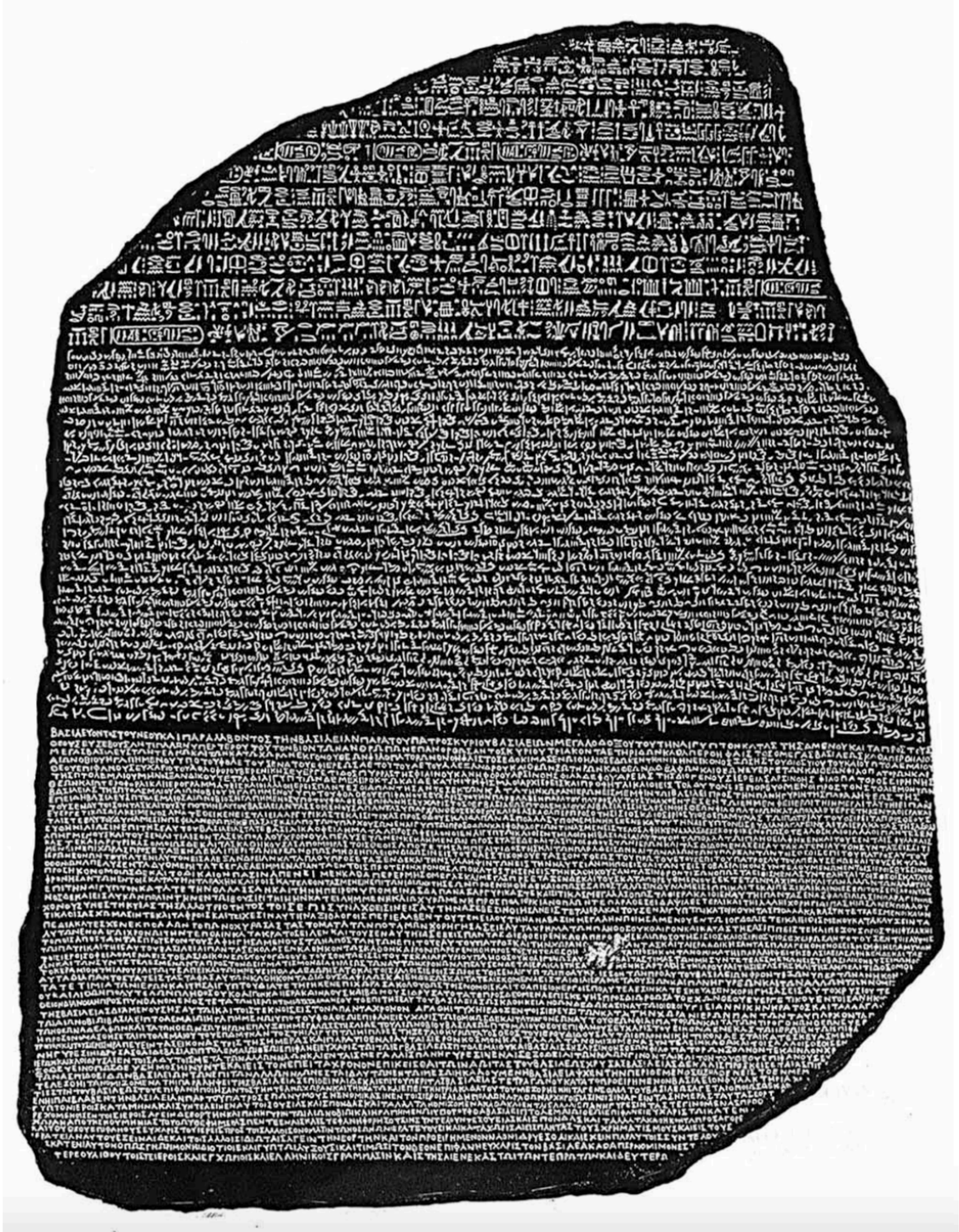
ASR

waveform → ***Automatic speech recognition*** → *text*



`Author of the…`



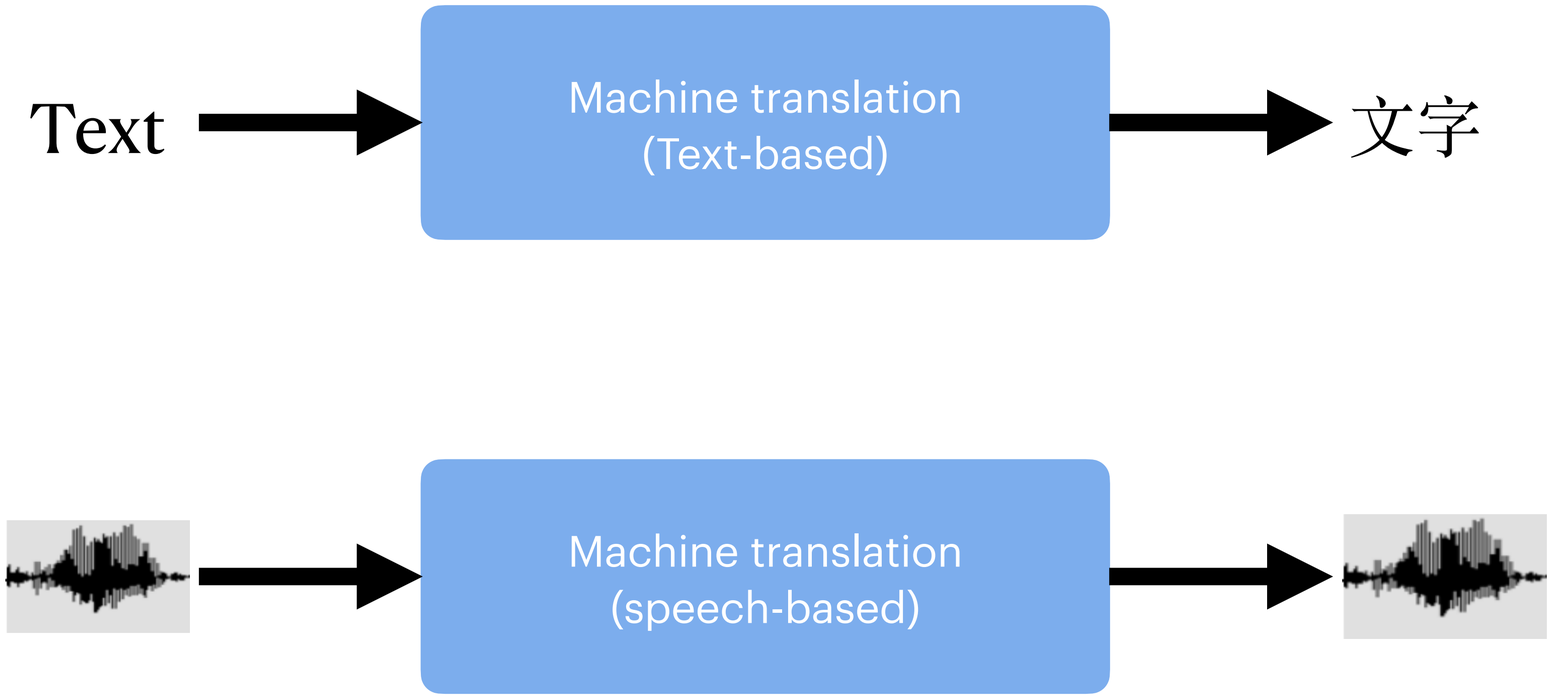


The Rosetta Stone

- The same text in three languages

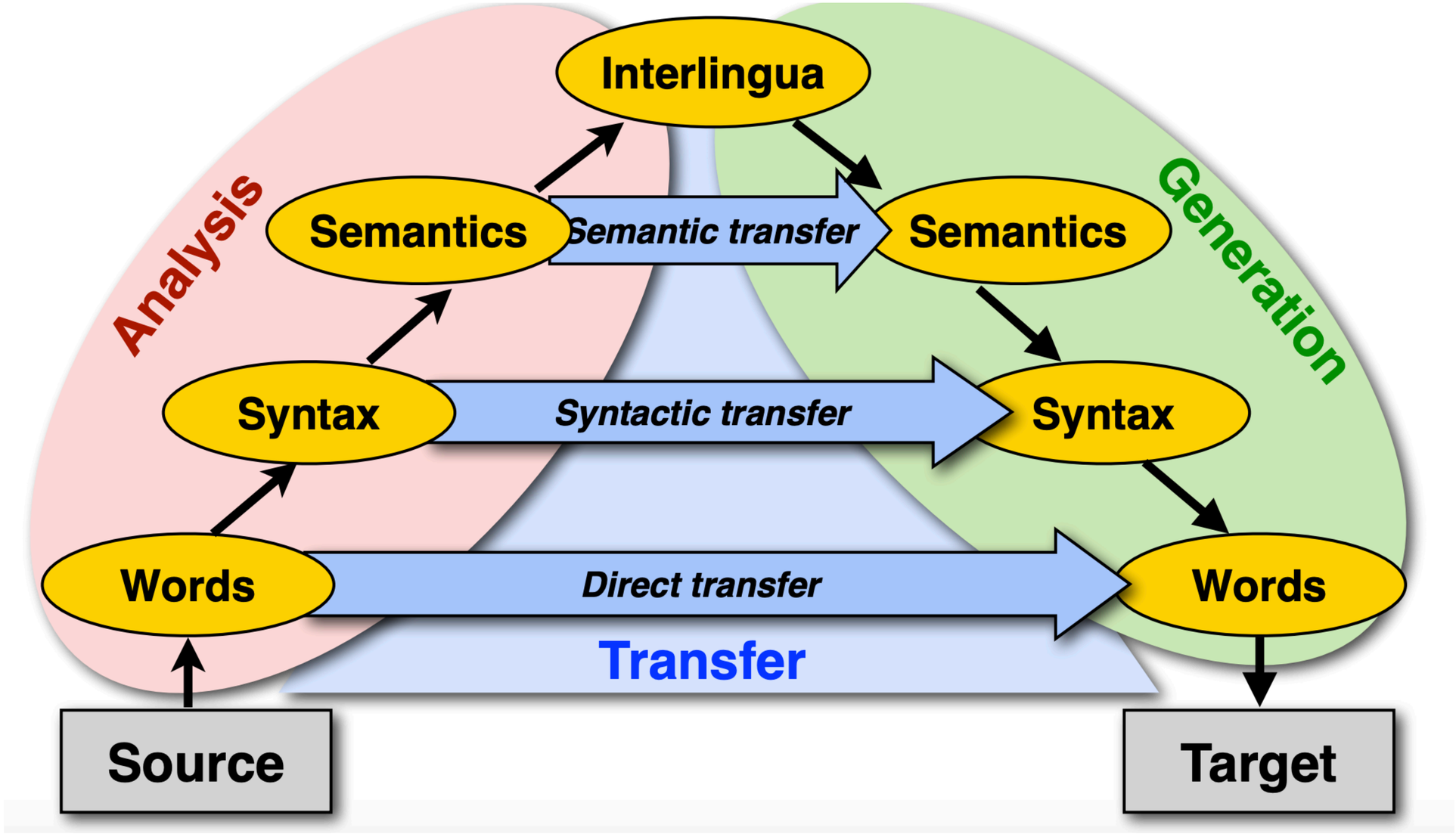
- This is an instance of parallel text
  - The Greek inscription allowed scholars to decipher the hieroglyphs

Machine Translation



Applications

The Vauquois triangle



https://www.youtube.com/watch?v=cgI_Zb1VMBE

Why machine translation is difficult?

- Lexical divergences
- Syntactic divergences
- Semantic divergences
- Word-to-Word correspondence divergences

Lexical divergences

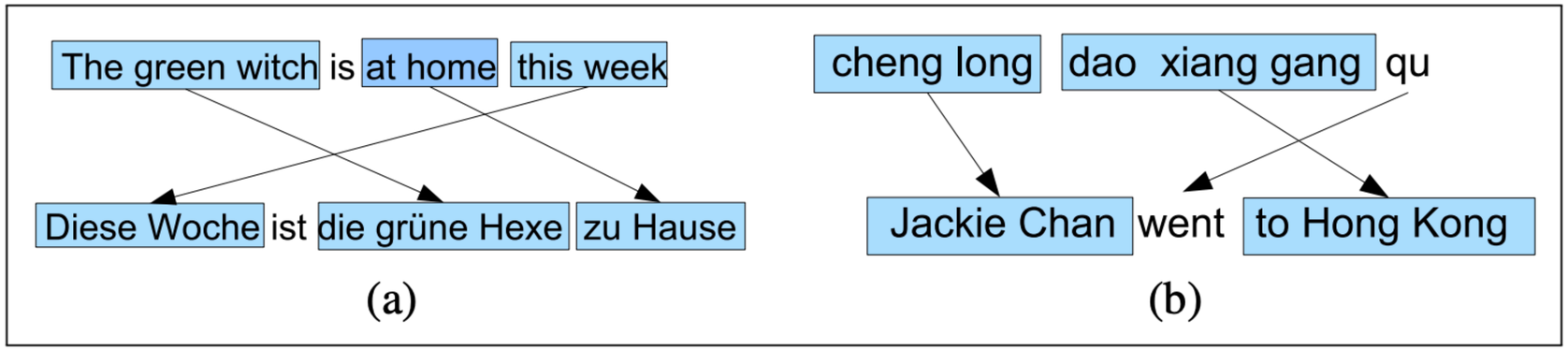
- English: brother vs Chinese: gege, didi (哥哥、弟弟)
- English: sister vs Chinese: jie jie, meimei (姐姐、妹妹)

- English: Play vs Chinese: 打、弹、...

- The different senses of homonymous words generally have different translations:
  - English-German:
    - (river) bank - Ufer
    - (financial) bank - Bank

Syntactic divergences

- Word order



- Head-marking vs. dependent-marking
  - Dependent-marking (English): the man's house
  - Head-marking (Hungarian): the man house-his

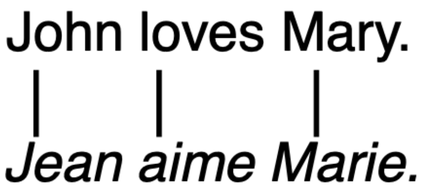
Semantic divergences

- English has a progressive aspect
  - ‘Peter swims’ vs. ‘Peter is swimming’

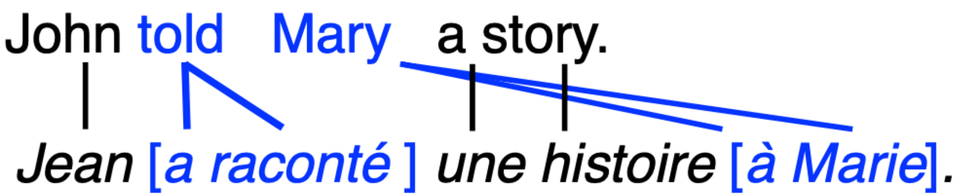
- German can only express this with an adverb
  - ‘Peter schwimmt’ vs. ‘Peter schwimmt gerade’ (‘swims currently’)

Word-to-word correspondences

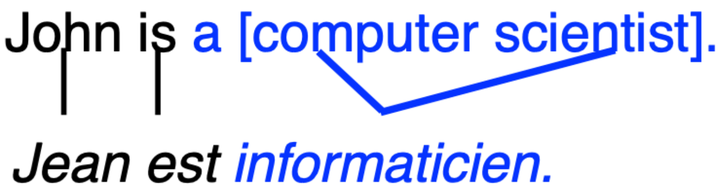
One to-one:



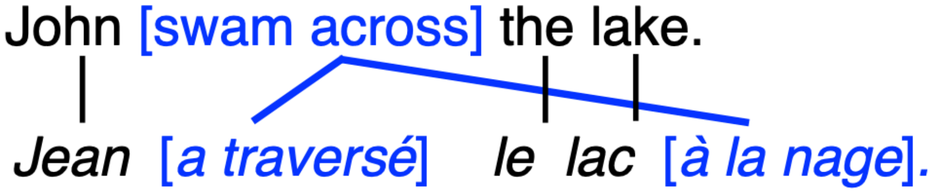
One-to-many:
(and reordering)



Many-to-one:
(and elision)



Many-to-many:



https://courses.grainger.illinois.edu/cs447/sp2023/Slides/Lecture16.pdf

Statistical MT

Given input in the source language, S,…

e.g. a Chinese sentence…

主席：各位議員，早晨。

… return the best translation in the target language, T^*

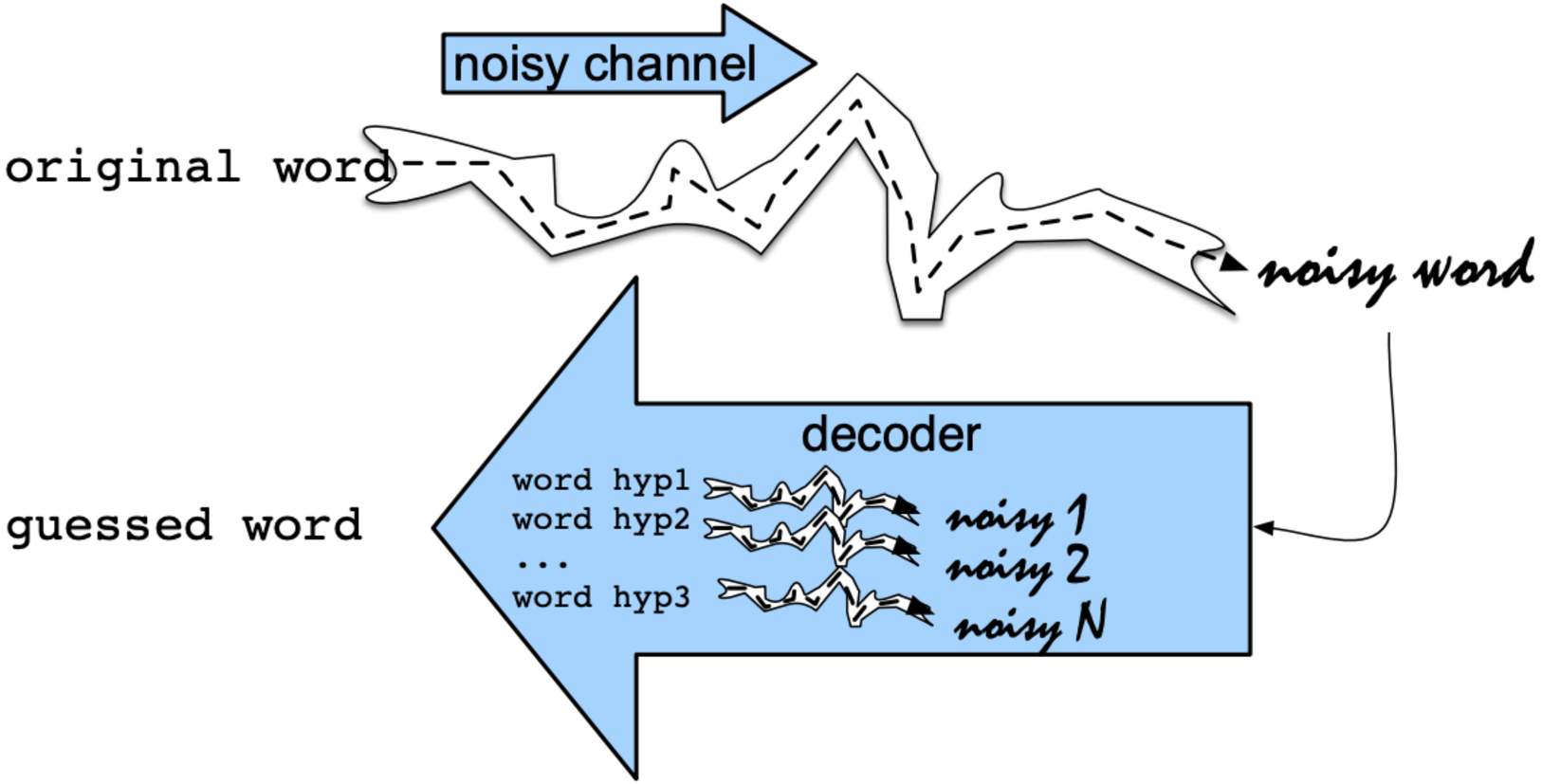
e.g in English:

President: Good morning, Honourable Members.

We can formalize this as $T^* = \text{argmax}_T P(T \mid S)$

https://courses.grainger.illinois.edu/cs447/sp2023/Slides/Lecture16.pdf

Noise channel model



https://web.stanford.edu/~jurafsky/slp3/B.pdf

Noise Channel model

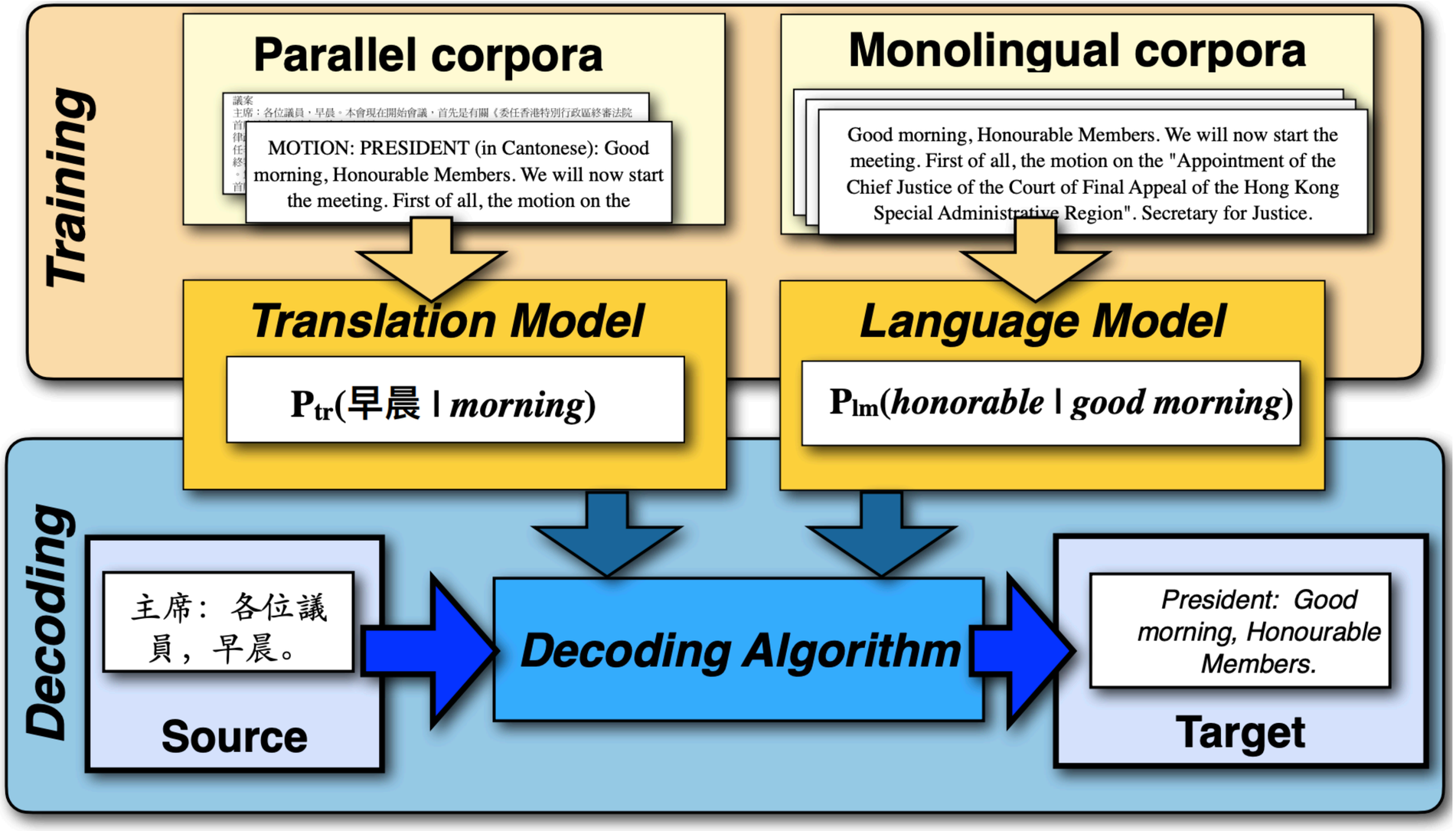
- Suppose we want to translate a foreign language to English, we can model $P(E|F)$

- by Bayes law, we have the equivalent equation

$$\hat{E} = \underset{E \in \text{ English}}{\operatorname{argmax}} \overbrace{P(F \mid E)}^{\text{translation model}} \overbrace{P(E)}^{\text{language model}}$$

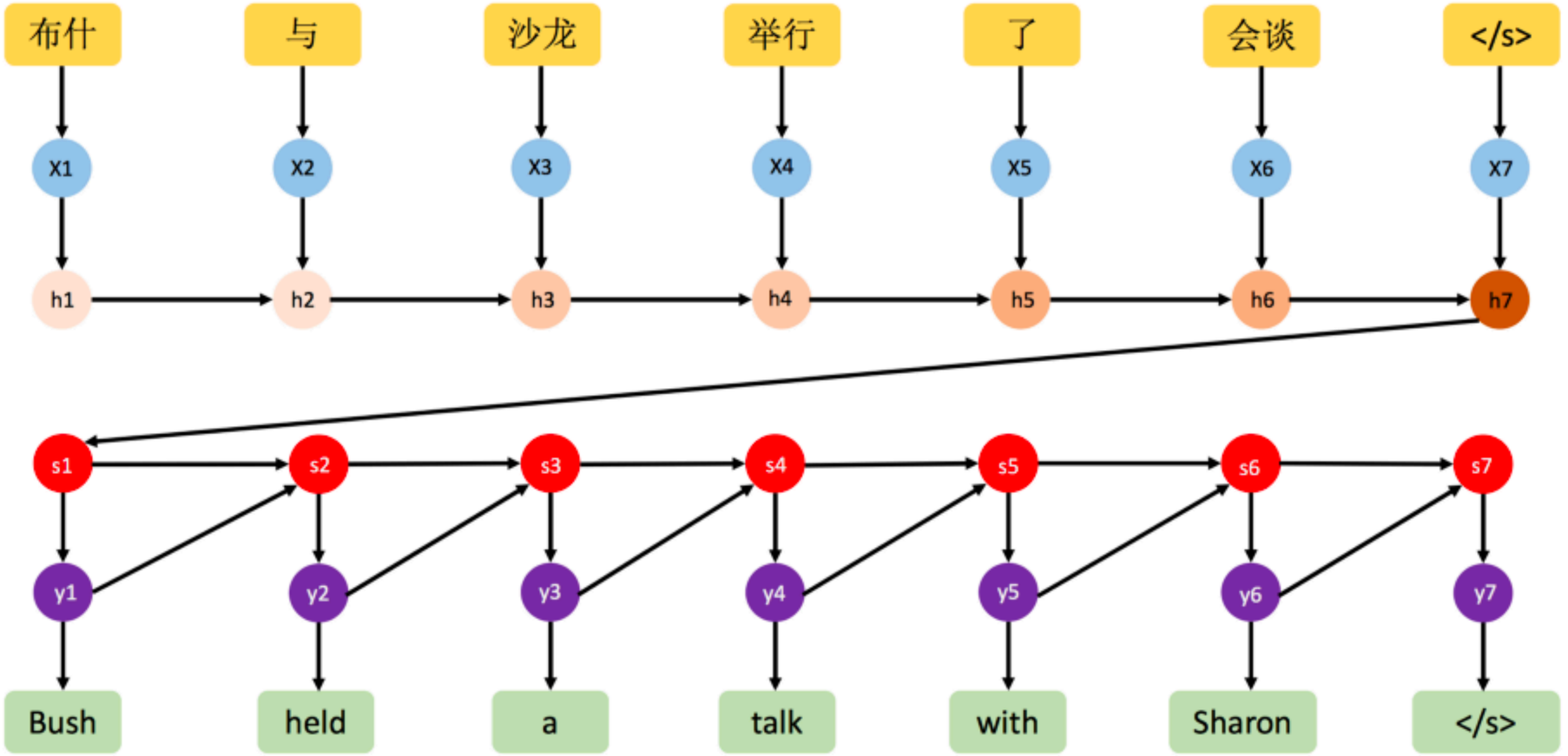
  - Intuitively, a good model of English, and a good English-to-foreign translator

Training and decoding



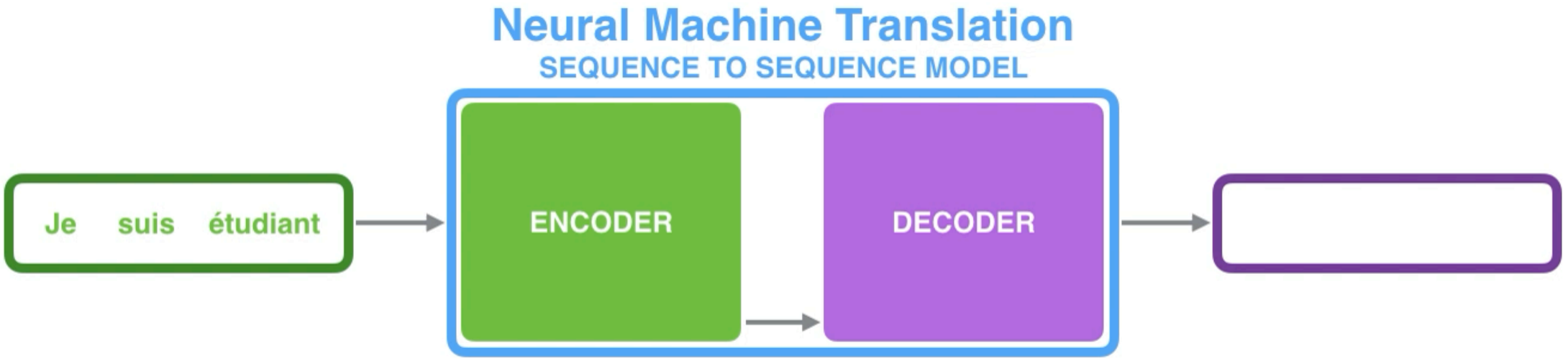
https://courses.grainger.illinois.edu/cs447/sp2023/Slides/Lecture16.pdf

Neural MT



(Sutskever et al., 2014)

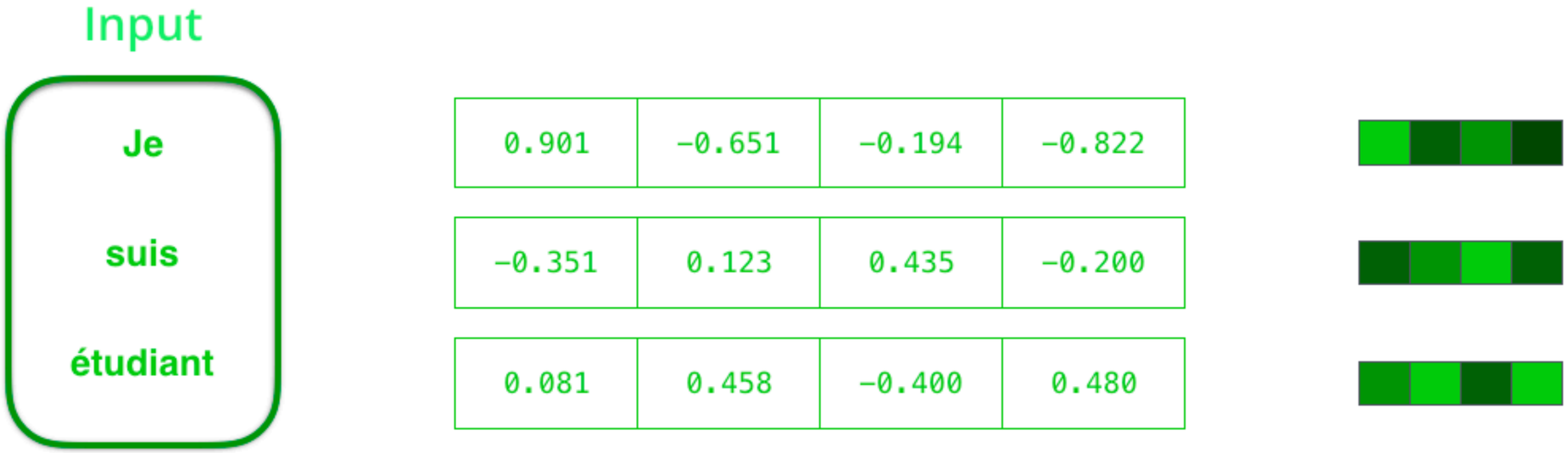
Encoder-decoder model



https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

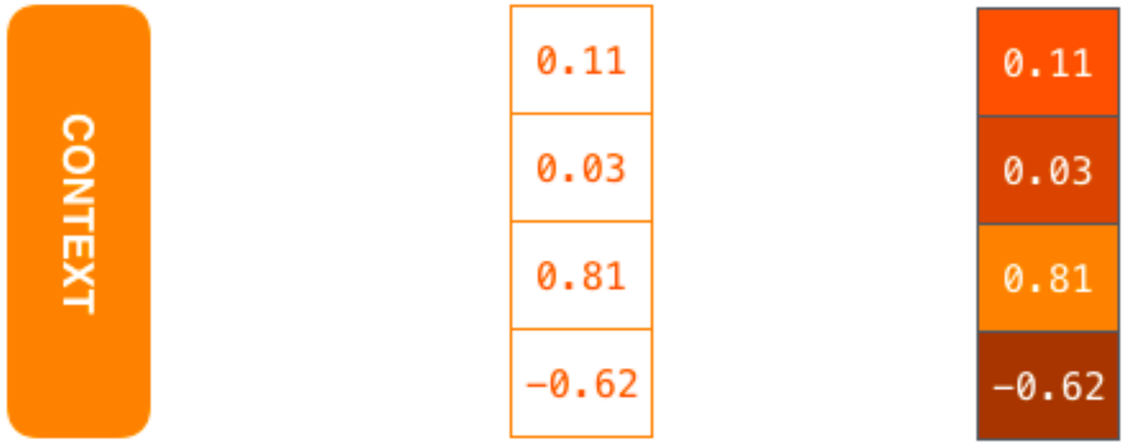
Input to the encoder

- Embedding

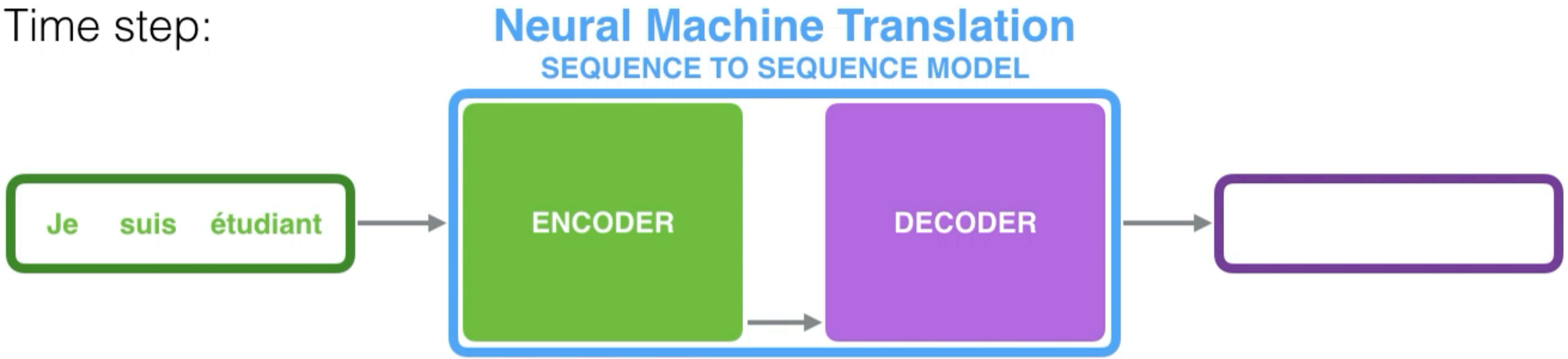


Context vector

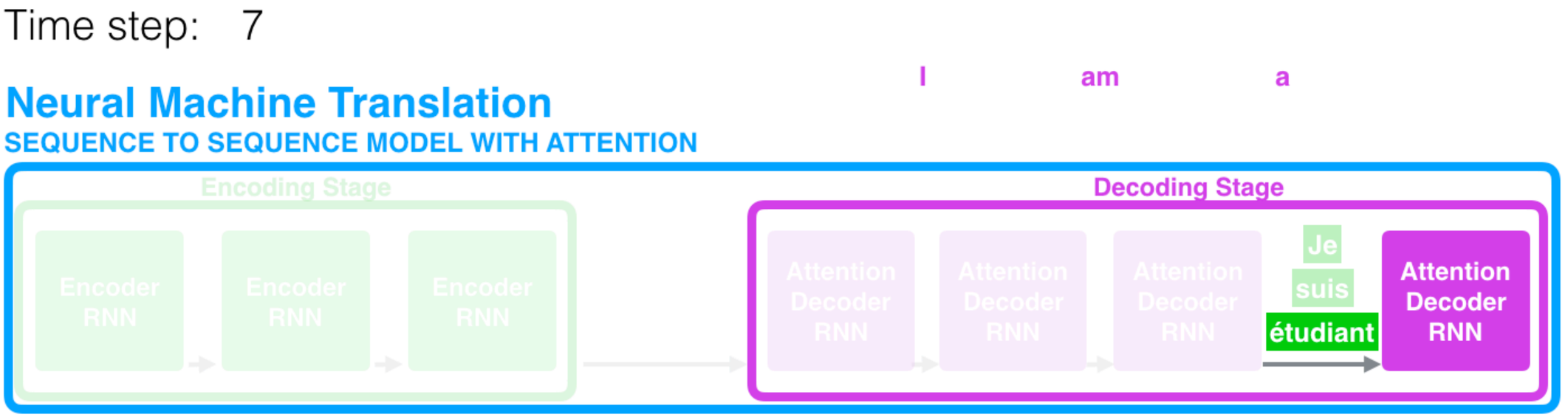
- A continuous vector from the encoder



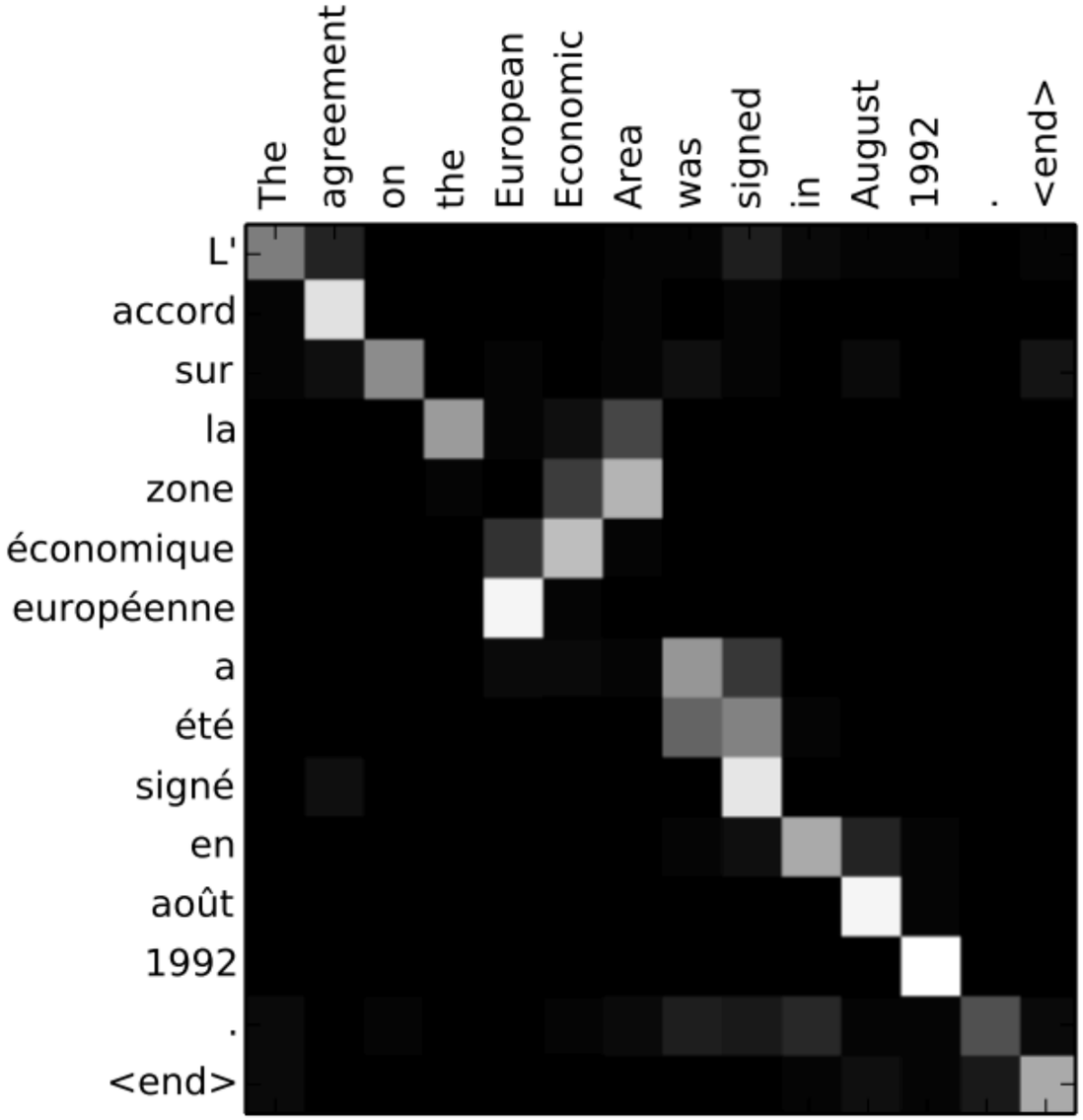
Seq2seq


Time step:
Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL
Je suis étudiant
ENCODER
DECODER

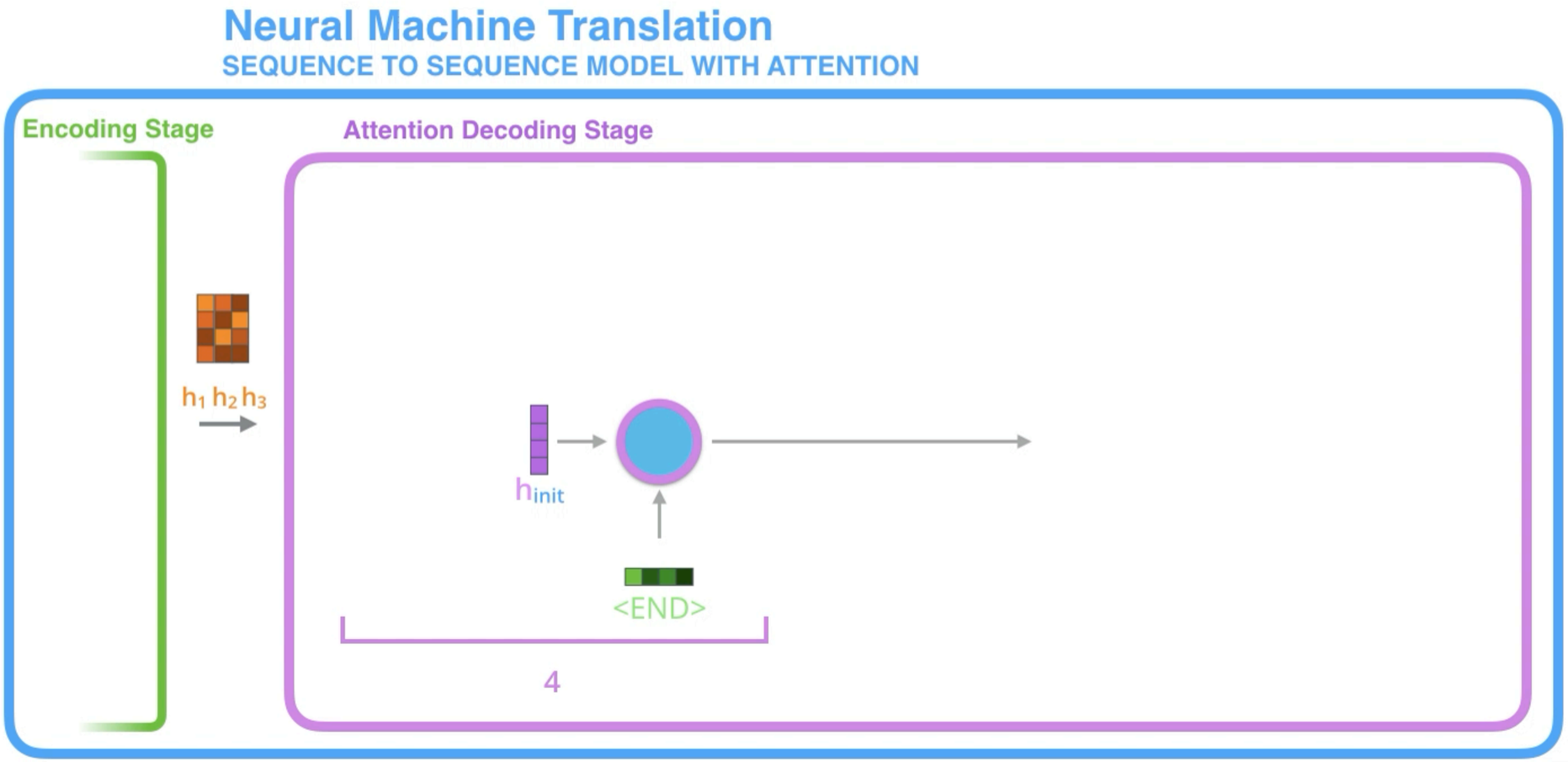
Let’s pay attention



Attention visualization

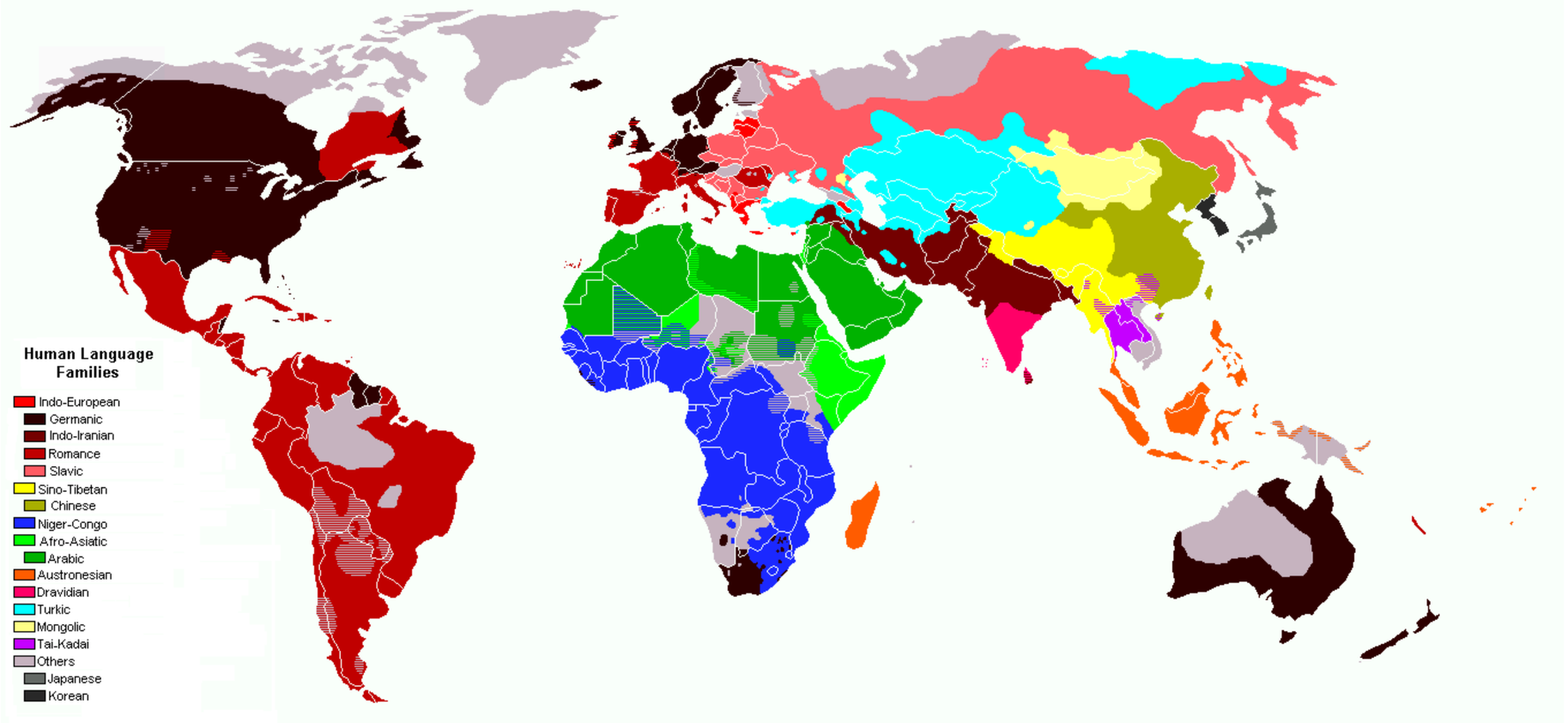


Seq2Seq with attention

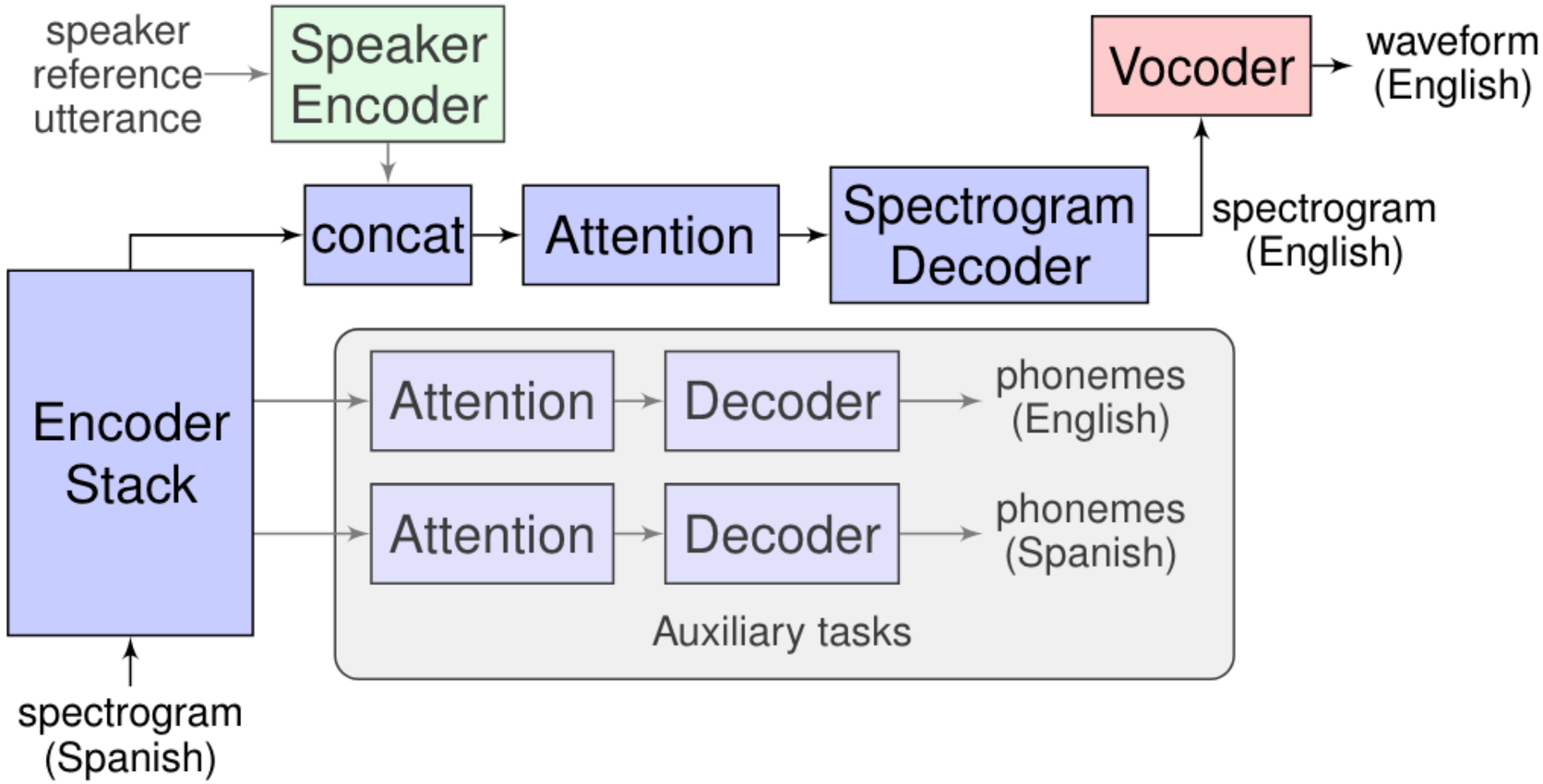


Speech-to-speech translation

- Some languages don’t have a written form

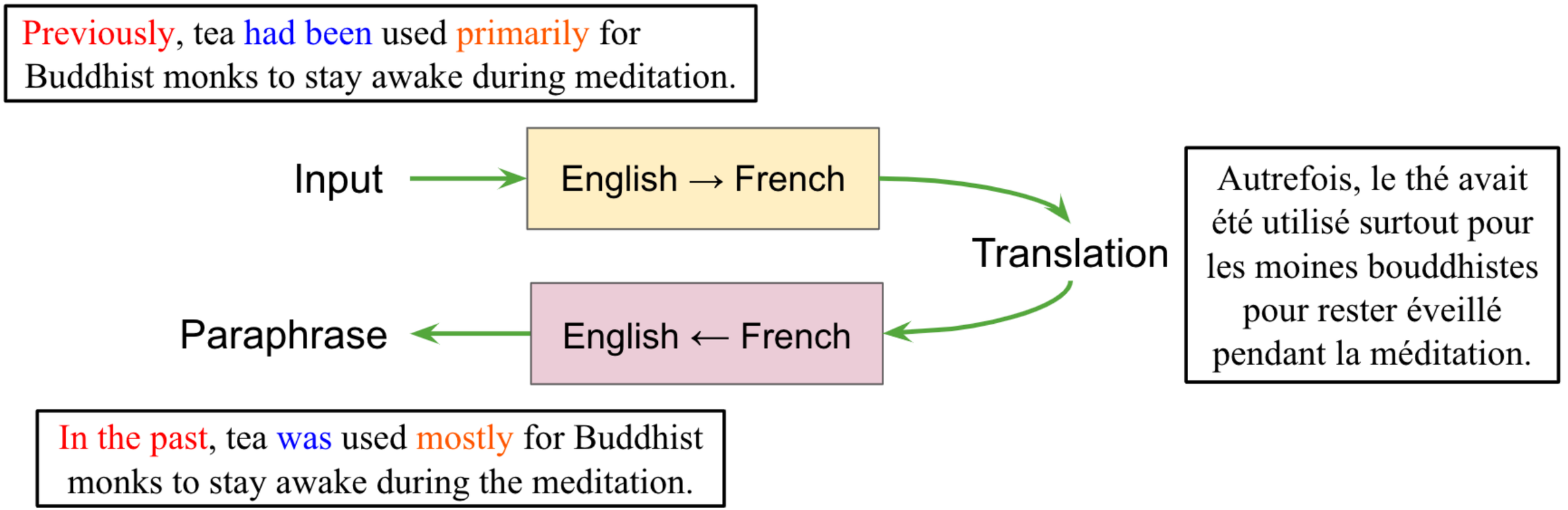


Speech-to-speech translation



Jia, Ye, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. "Direct speech-to-speech translation with a sequence-to-sequence model." *arXiv preprint arXiv:1904.06037* (2019).

Data augmentation: back translation



Summary

- Lexical, syntactic, semantic divergences make MT difficult
- Statistical machine translation and noise channel model
- End-to-end neural machine translation
- Speech-to-speech translation for languages without written form

Reading

- Chapter 10: Machine translation
  - https://web.stanford.edu/~jurafsky/slp3/13.pdf