



MACHINE TRANSLATION

Luísa Coheur

Overview

- Learning objectives
- Topics
 - Motivation
 - The MT task
 - Approaches
 - Classic MT
 - Statistical-based MT
 - Deep Learning-based MT
 - MT shared tasks, datasets and metrics
 - Translation to sign languages
- Key takeaways
- Suggested readings

LEARNING OBJECTIVES

LEARNING OBJECTIVES

- Focus on a specific NLP task – Machine Translation – and understand how it evolved through the years
- Relate the MT ingredients with the ones we have studied (and, revise, in this way the metrics, techniques, architectures, etc., we have studied in this course)

TOPICS

Overview

- Learning objectives
- Topics
 - Motivation
 - The MT task
 - Approaches
 - Classic MT
 - Statistical-based MT
 - Deep Learning-based MT
 - MT shared tasks, datasets and metrics
 - Translation to sign languages
- Key takeaways
- Suggested readings

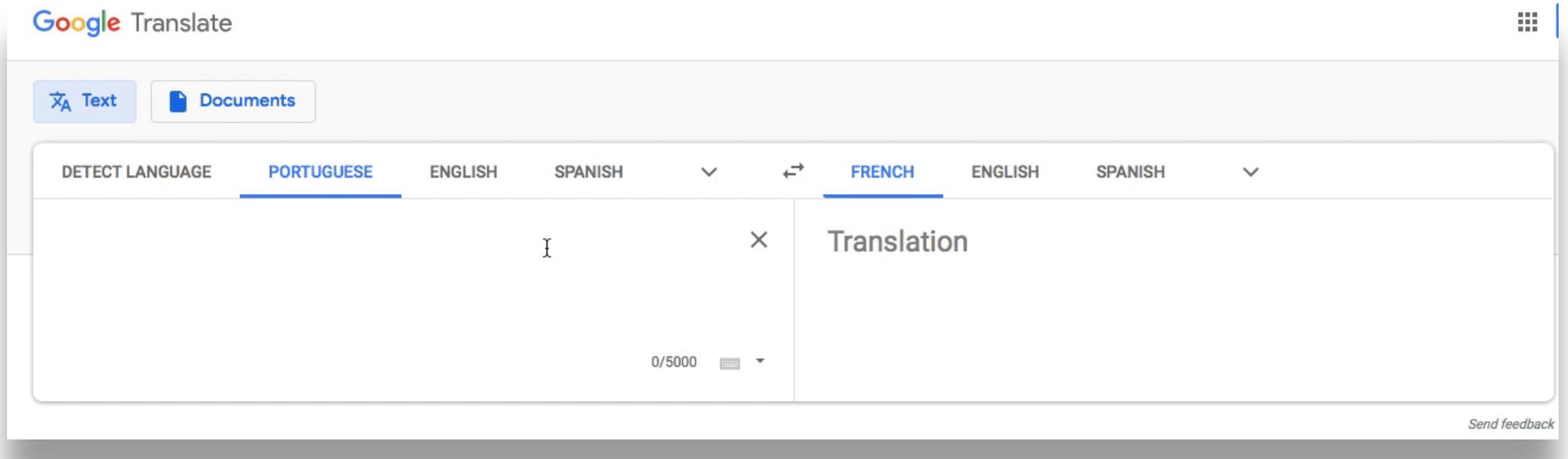
MOTIVATION

- The dream: translation allows to eliminate language barriers



MOTIVATION

- Machine Translation (MT) is part of our lives



MOTIVATION

- MT is costly:
 - Example (taken from [source](#)):
 - The workload of Commission translators went from about 2 million pages in 2013 to 2.5 million in 2022
 - The European Commission's spending on translation increased from €26.5 million in 2012 to €35.8 million in 2023



MOTIVATION

- MT is not simple:
 - One word in one language can be translated in many words in other language; sometimes there is no obvious/exact translation

MOTIVATION

- Example:
 - Doudou (FR) =? boneco de estimação



MOTIVATION

- Example:
 - Eu sou/estou vs. I'm/Je suis

MOTIVATION

- **Collocations** are especially tricky
 - Not many learners know they exist
 - There is no apparent reasoning or rule behind them
 - If you asked a native speaker why they use certain collocations, they won't be able to give you any justification besides "I just do", or "It sounds rights"

MOTIVATION

What is a
collocation?



MOTIVATION

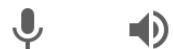
- Collocations are a combination of words in a language that happens very often and more frequently than would happen by chance. Using a different combination of words sounds unnatural
- Example (collocations):
 - “I missed the flight” vs. “perdi o avião” (something like “I lost the plane”)
 - “The employee cleared the table.” vs “O empregado levantou a mesa” (something like “The employee raised the table”).

MOTIVATION

O empregado levantou
a mesa do jantar

x

The waiter raised the
dinner table



Feedback

O empregado levantou
a mesa do jantar e
levou o bolo

x

The waiter cleared
the dinner table and
took the cake



Feedback

MOTIVATION



Em português, a frase "O empregado levantou a mesa" pode ser interpretada de diferentes maneiras dependendo do que você quer dizer.

1. Se você quer dizer que o empregado ergueu fisicamente a mesa (por exemplo, para movê-la), a frase está correta como "O empregado levantou a mesa."
2. No entanto, se a intenção é dizer que o empregado limpou ou recolheu os objetos da mesa após uma refeição, o mais comum seria dizer "O empregado tirou a mesa" ou "O empregado recolheu a mesa."

Essas expressões variam conforme o contexto e o que se quer comunicar.

Não é verdade. Em Português, dizemos "levantar a mesa" para dizer que recolheu os objectos.



Você está correto! Em algumas variantes do português, como no português europeu, "levantar a mesa" é uma expressão idiomática que significa justamente recolher os objetos após uma refeição, como pratos, talheres e copos. Agradeço a correção!

MOTIVATION

- MT is not simple:
 - Structural Divergences: differences in word order are common even between similar languages



"Precisely, cousin Asterix. We need the magic potion to fight the Roman armies."

MOTIVATION

- MT is not simple (we have seen this):
 - bias problem

LC

You

Translate to Portuguese: the doctor said "hello" to the nurse

< 2 / 2 >



ChatGPT

O médico disse "olá" para a enfermeira.

Overview

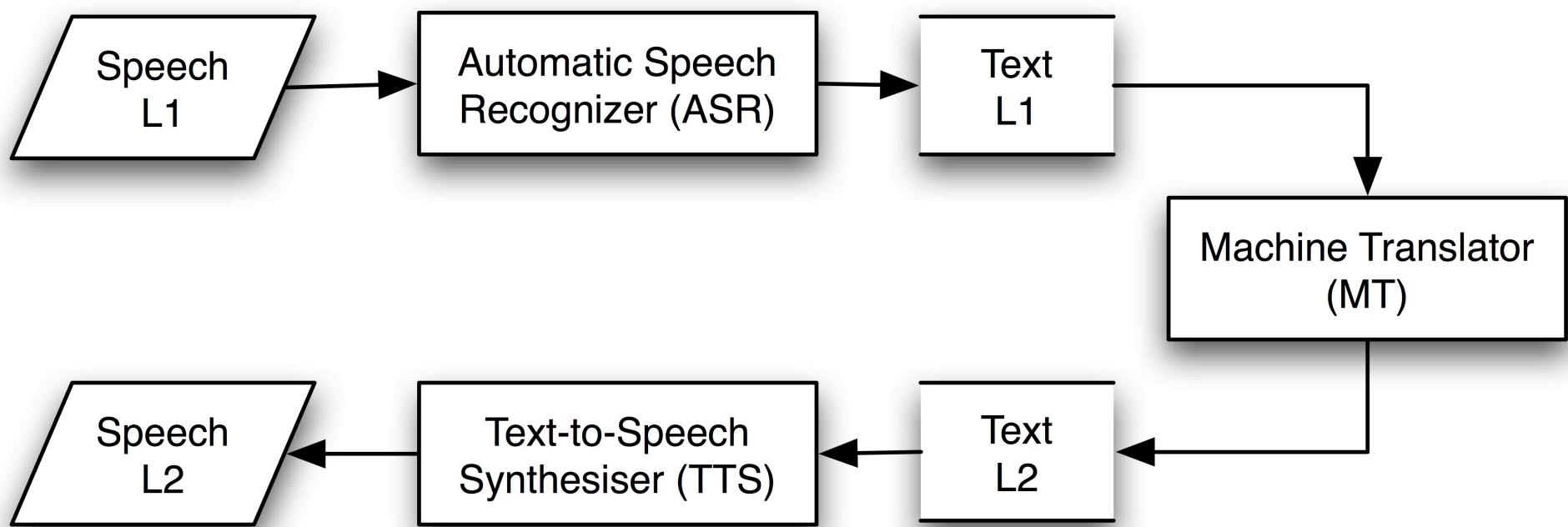
- Learning objectives
- Topics
 - Motivation
 - [The MT task](#)
 - Approaches
 - Classic MT
 - Statistical-based MT
 - Deep Learning-based MT
 - MT shared tasks, datasets and metrics
 - Translation to sign languages
- Key takeaways
- Suggested readings

THE MT TASK

- MT is the use of computers to automate some or all of the process of translating from one language to another (Jurafsky)

THE MT TASK

- Do not forget speech to speech MT
 - which might include Voice Morphing (process of transforming or modifying a person's voice to sound like someone else's or to achieve a desired effect)



OLD DEMO (around 2012)



THE MT TASK

- Do not forget sign languages
 - next

Overview

- Learning objectives
- Topics
 - Motivation
 - The MT task
 - Approaches
 - [Classic MT](#)
 - Statistical-based MT
 - Deep Learning-based MT
 - MT shared tasks, datasets and metrics
 - Translation to sign languages
- Key takeaways
- Suggested readings

CLASSIC APPROACH

(From Jurafsky)

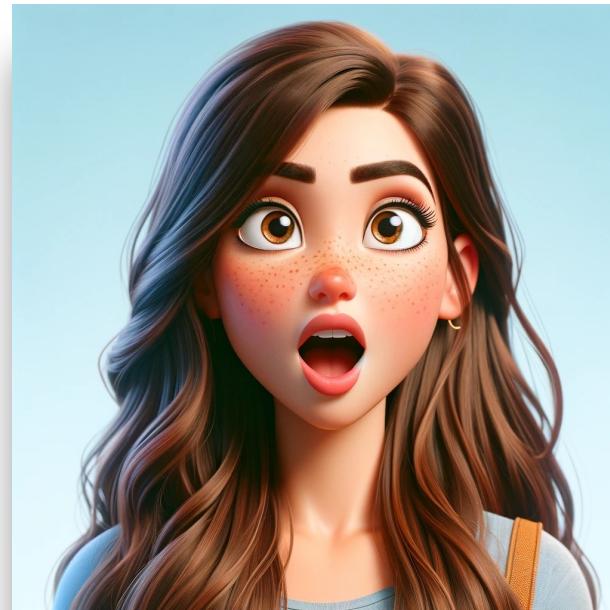


- 1946: Booth and Weaver discuss MT at Rockefeller foundation in New York
- 1947-48: idea of dictionary-based direct translation
- 1949: Warren Weaver
 - Idea: we have methods for solving cryptographic problems. So, maybe, we already have methods for solving translation...
 - it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code."

CLASSIC APPROACH

(From Jurafsky)

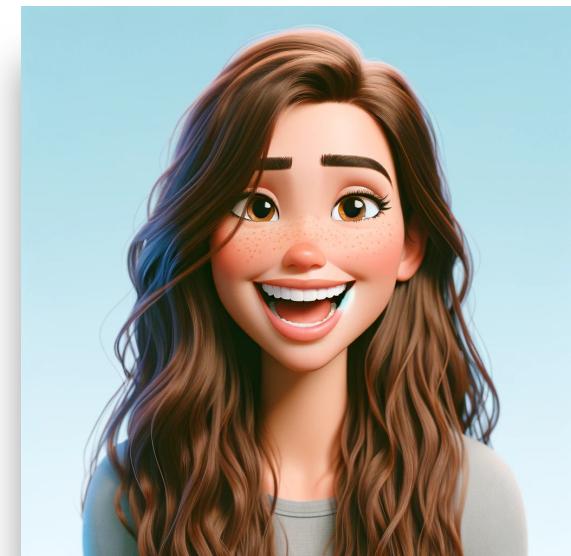
- 1952: all 18 MT researchers in world meet at MIT



CLASSIC APPROACH

(From Jurafsky)

- 1954: IBM/Georgetown Demo Russian-English MT
 - Six grammar rules and 250 lexical items in its vocabulary
 - The experiment did encourage governments to invest in computational linguistics
 - Authors claimed that within 3-5 years, MT could well be a solved problem
- 1955-65: lots of labs take up MT



DIRECT TRANSLATION

- From Reddit
 - French: pomme de terre
 - English: potato
 - Literal English translation: apple of the earth
- French: quatre-vingt-dix
 - English: ninety
 - Literal English translation: four-twenty-ten
- French: Qu'est-ce que c'est que ça?
 - English: What is that?
 - Literal: What is it that it is that that?

ACTIVE LEARNING MOMENT



EXERCISE

Translate to Portuguese (or your own native language),
word by word:

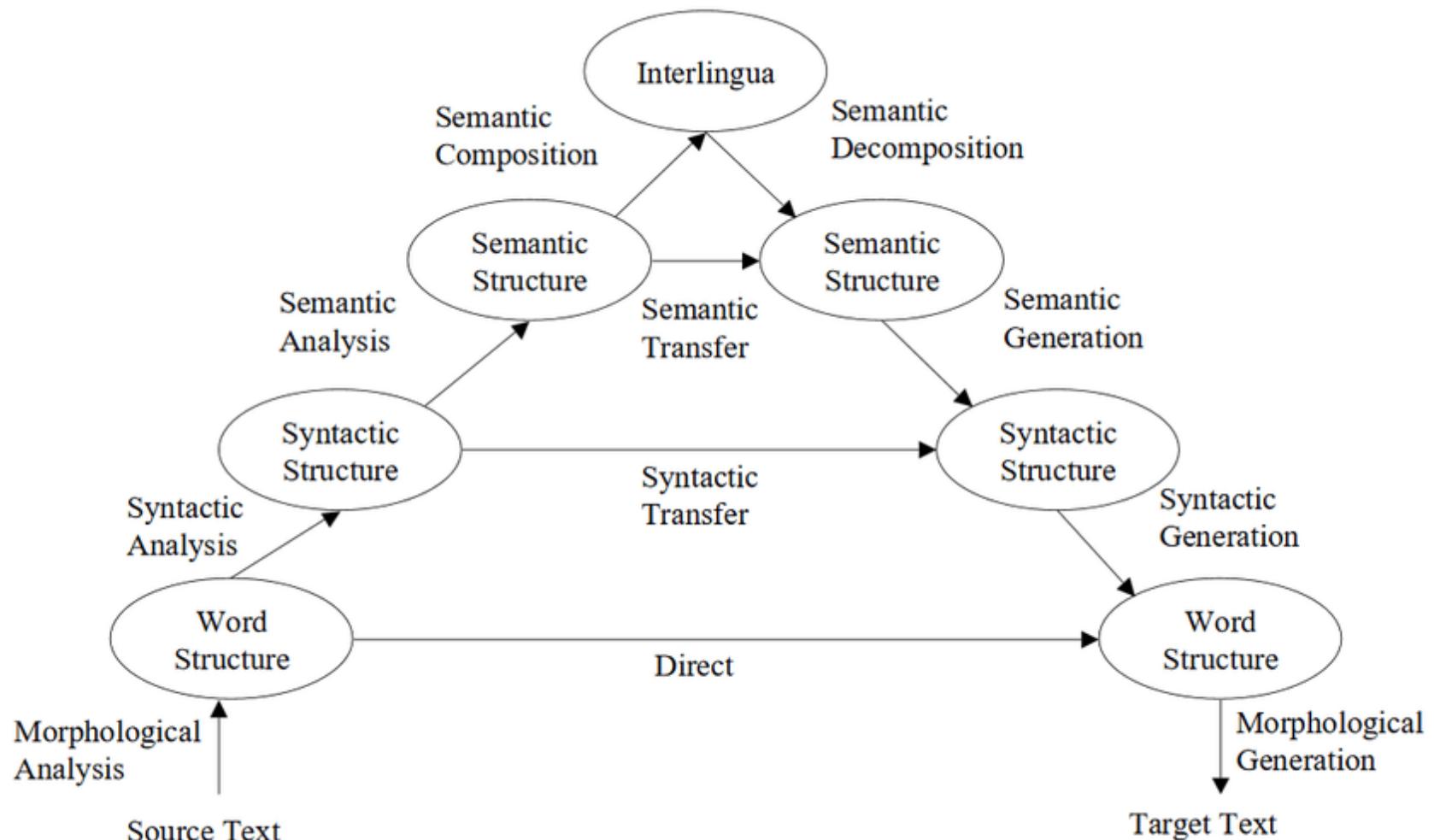
“I am looking forward to meeting my new boss.”

Eu estou olhando (em) frente para encontrar meu novo
chefe.

Correct: Estou ansiosa para conhecer meu novo chefe.

CLASSIC APPROACH

- THE VAUQUOIS TRIANGLE (1968)



INTERLINGUA

- In interlingua approaches, the source language text is mapped into some abstract **meaning representation** called an **interlingua**. The target language is generated from this interlingua representation
 - Example: map Portuguese into a First Order Logic (FOL) formula and then, the formula in FOL into English

Overview

- Learning objectives
- Topics
 - Motivation
 - The MT task
 - Approaches
 - Classic MT
 - Statistical-based MT
 - Deep Learning-based MT
 - MT shared tasks, datasets and metrics
 - Translation to sign languages
- Key takeaways
- Suggested readings

ACTIVE LEARNING MOMENT



EXERCISE (from [Knight 97])

Consider the following parallel corpus, with translations from Centauri to Arcturan:

- 1a. erok sprok izok hihok ghirok .
- 1b. totat dat arrat vat hilat .
- 2a. ok-voon anok drok brok jok .
- 2b. at-voon krat pippat sat lat .
- 3a. wiwok farok izok stok .
- 3b. totat jjat quat cat .
- 4a. lalok farok ororok lalok sprok izok enemok .
- 4b. wat jjat bichat wat dat vat neat.
- 5a. lalok brok anok plok nok .
- 5b. iat lat pippat rrat nnat .

Translate to Arcturan:

farok crrrok hihok yorok clok kantok ok-yurp

EXERCISE (from [Knight 97])

Consider the following parallel corpus, with translations from Centauri to Arcturan:

- 1a. erok sprok izok hihok ghirok .
- 1b. totat dat arrat vat hilat .
- 2a. ok-voon anok drok brok jok .
- 2b. at-voon krat pippat sat lat .
- 3a. wiwok **farok** izok stok .
- 3b. totat jjat quat cat .
- 4a. lalok **farok** ororok lalok sprok izok enemok .
- 4b. wat jjat bichat wat dat vat neat.
- 5a. lalok brok anok plok nok .
- 5b. iat lat pippat rrat nnat .

Translate to Arcturan:

farok crrrok hihok yorok clok kantok ok-yurp

EXERCISE (from [Knight 97])

Consider the following parallel corpus, with translations from Centauri to Arcturan:

- 1a. erok sprok izok hihok ghirok .
- 1b. totat dat arrat vat hilat .
- 2a. ok-voon anok drok brok jok .
- 2b. at-voon krat pippat sat lat .
- 3a. wiwok **farok** izok stok .
- 3b. totat **jjat** quat cat .
- 4a. lalok **farok** ororok lalok sprok izok enemok .
- 4b. wat **jjat** bichat wat dat vat neat.
- 5a. lalok brok anok plok nok .
- 5b. iat lat pippat rrat nnat .

Translate to Arctuan:

farok crrrok hihok yorok clok kantok ok-yurp

STATISTICAL MACHINE TRANSLATION

- Since the 1990s, with the emergence of large amounts of data that could be used for training, there has been significant development of systems based on statistical methods
- A very important dataset was EuroParl



STATISTICAL MACHINE TRANSLATION

- Statistical Machine Translation (SMT) is the name for a class of approaches that build probabilistic models of **faithfulness** (or fidelity or adequacy) and **fluency**, and then **combine these models to choose the most probable translation**

STATISTICAL MACHINE TRANSLATION

Being S a source language sentence, SMT models its translation to a target sentence T' as:

$$T' = \operatorname{argmax}_T P(T, S) P(T)$$

Translation Model
(faithfulness)

Language Model
(fluency)

$S = \text{Estou cansado}$	Faithfulness	Fluency
I'm exhausted	3	5
Tired me	5	2
I love cookies	0	5



STATISTICAL MACHINE TRANSLATION

- How does it work?
 - Various translation hypotheses are created
 - A probability is assigned to each hypothesis
 - The most likely hypothesis is chosen

STATISTICAL MACHINE TRANSLATION

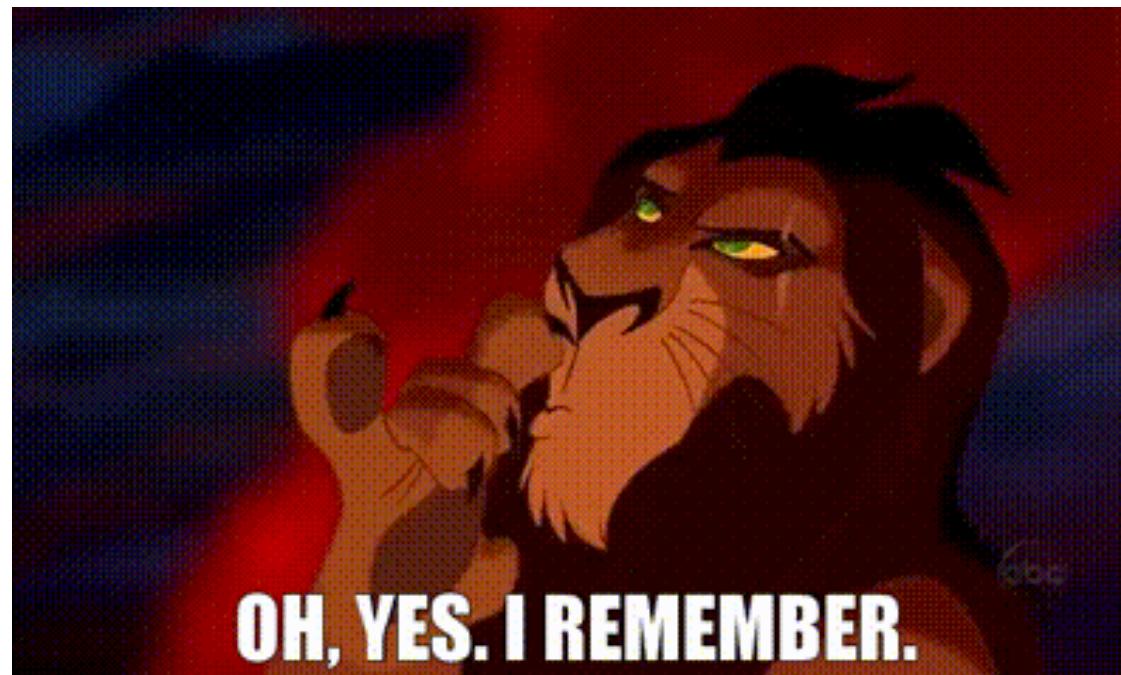
- Big boost in MT
 - Good results can be quickly achieved
 - Can be used for any language pair, as long as training data is available,
 - But (there is always a but) many data are required, and the models created are highly dependent on the data

Overview

- Learning objectives
- Topics
 - Motivation
 - The MT task
 - Approaches
 - Classic MT
 - Statistical-based MT
 - Deep Learning-based MT
 - MT shared tasks, datasets and metrics
 - Translation to sign languages
- Key takeaways
- Suggested readings

DEEP LEARNING-BASED MACHINE TRANSLATION

- RNNs
- Seq2Seq
- Attention
- Transformers



<https://trailers.getyarn.io/yarn-clip/81e73630-d6c1-4bce-a05d-654f36c56f2e/gif>

DEEP LEARNING-BASED MACHINE TRANSLATION

- Multilingual language models are trained on data from multiple languages simultaneously, allowing them to learn representations that capture commonalities and differences across languages
 - The model learns a shared vocabulary and tokenization strategy across multiple languages. Monolingual text from each language is used to create subword units or tokens that can handle the diversity of languages

DEEP LEARNING-BASED MACHINE TRANSLATION

- Multilingual models alleviate the dependency on large parallel corpora by using monolingual data and leveraging the transfer of linguistic structures learned from well-resourced languages
 - Monolingual Data:
 - used primarily the initial stages of training for learning language-specific features
 - Parallel Data:
 - used particularly in fine-tuning stages

DEEP LEARNING-BASED MACHINE TRANSLATION

- As the model trains on many languages at once, it learns to place semantically similar words close together
 - Example
 - “cat” (English), “gato” (Spanish), “chat” (French)
 - During training, the model sees these words used in similar contexts:
 - “The __ is sleeping.”
 - “El __ duerme.”
 - “Le __ dort.”
 - The model learns that these tokens often appear
 - Near words about animals, homes, sleep, etc.
 - In similar sentence positions (subject position).
 - As a result, it adjusts their vector representations so that “cat,” “gato,” and “chat” end up close together in the embedding space

DEEP LEARNING-BASED MACHINE TRANSLATION

- The next step (with parallel data) fine-tunes the model's cross-lingual mapping — helping it understand which words or phrases correspond across languages.
 - Example:
 - The model learns that “cat” in English corresponds to “gato” in Spanish through a parallel dataset.

Overview

- Learning objectives
- Topics
 - Motivation
 - The MT task
 - Approaches
 - Classic MT
 - Statistical-based MT
 - Deep Learning-based MT
 - [MT shared tasks, datasets and metrics](#)
 - Translation to sign languages
- Key takeaways
- Suggested readings

MT SHARED TASKS, DATASETS AND METRICS

- Workshop on Machine Translation (WST) since 2006 (now it is a conference)
 - Some current shared tasks
 - General machine translation task (former News task)
 - Biomedical translation task
 - Multimodal translation task
 - Sign language translation task
 - ...
 - Quality estimation task
 - Some previous shared tasks
 - Medical text translation task
 - ...

MT SHARED TASKS, DATASETS AND METRICS

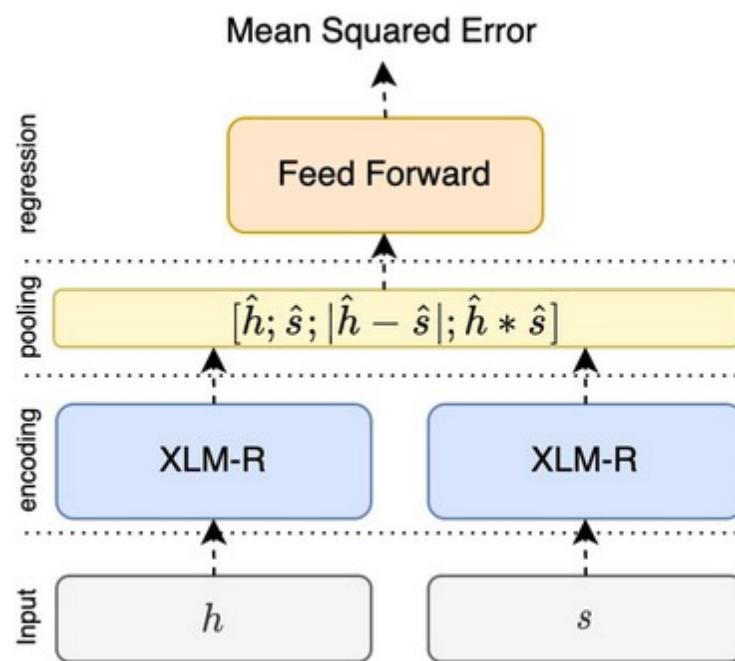
- Some Datasets:
 - Datasets from the WMT shared tasks
 - Europarl
 - Europarl-ST (spoken translation)
 - Open subtitles (check OPUS page)
 - MLQA (Multilingual Question/Answering)
 - ...

MT SHARED TASKS, DATASETS AND METRICS

- Some metrics:
 - BLUE
 - METEOR
 - TER (Translation Edit Rate or Translation Error Rate)
 - ...

MT SHARED TASKS, DATASETS AND METRICS

- Some metrics (cont.):
 - COMET (Crosslingual Optimized Metric for Evaluation of Translation) combines multilingual pre-trained models and [human quality assessments](#) to predict translation quality



Generated by DALL-E



KEY TAKEAWAYS

KEY TAKEAWAYS

- MT concepts (translation vs. language model, BLEU, COMET, ...)
- The MT task is a mirror of the NLP field
- It had a rule-base phase, a statistical/classical machine learning phase and now it is in the deep learning phase
- It has its own metrics, shared tasks, conferences/workshops

SUGGESTED READINGS

SUGGESTED READINGS

- Jurafsky, chapter 25
 - sections 25.2 to 25.9, inclusive
 - A Statistical MT tutorial Workbook, Kevin Knight, 1999