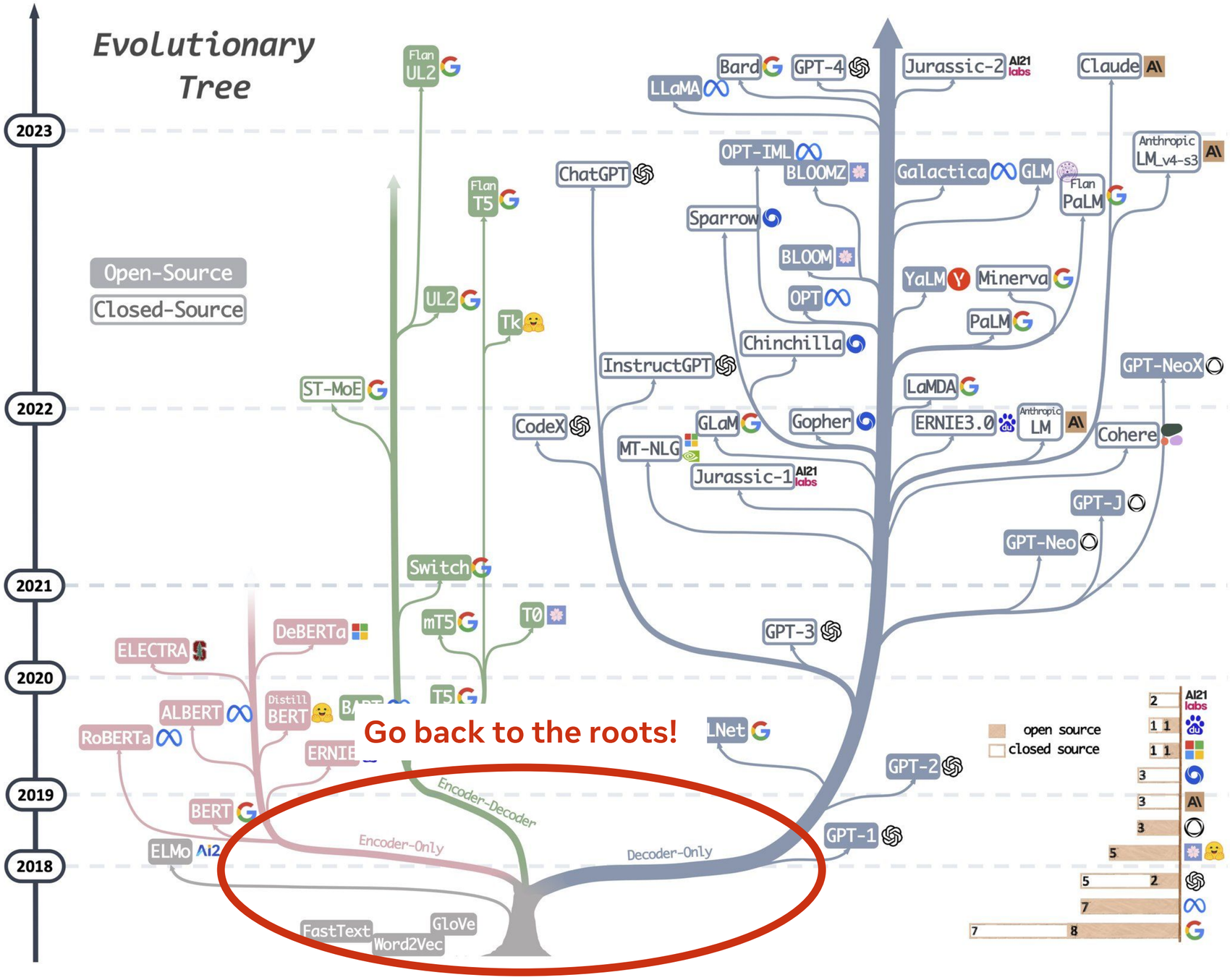
The State of Multilingual and Multimodal NLP

Maha Elbayad
Research Scientist, FAIR (Meta AI)

ThinkAI, May 7th 2023

Meta AI

LLMs


Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Flan UL2
Flan T5
UL2
Tk
ST-MoE
Switch
mT5
T0
T5
ELECTRA
DeBERTa
ALBERT
Distill BERT
RoBERTa
BERT
ELMo
Bard
GPT-4
LLaMA
Jurassic-2
Claude
OPT-IML
BLOOMZ
ChatGPT
Galactica
GLM
Flan PaLM
Anthropic LM_v4-s3
Sparrow
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
LaMDA
GPT-NeoX
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
GPT-3
LNet
GPT-2
GPT-1
Go back to the roots!
Encoder-Decoder
Encoder-Only
Decoder-Only
FastText
GloVe
Word2Vec
open source
closed source

Language Models 101

1. How to **represent text**?
2. What is a **Language model**?
3. What is a **Conditional Language Model**?

A basic setup

Computer vision: Binary image classification.

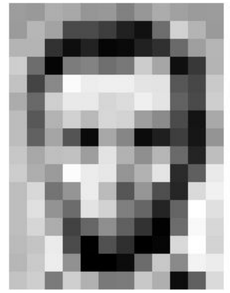
Vector of pixel values

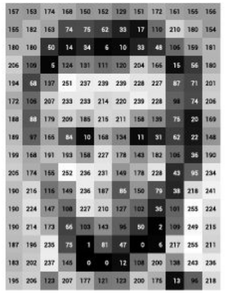
x

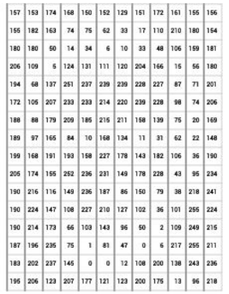
Model f_θ

$\hat{y}$ = dog?

0 1







https://ai.stanford.edu/~syyeung/cvweb/tutorial1.html

NLP: Sentiment prediction of movie reviews.

There is never a dull moment in this movie. Wonderful visuals and good actors.

?

x

Model f_θ

$\hat{y}$ = positive review?

0 1

How to represent this input?

The focus in this 101 will be on **representation learning.** We assume that:

1. we can evaluate a loss function that measures the error of our model on some training data.
2. we know how to optimize this loss function wrt the model parameters θ.

Text representation

Given a vocabulary $\mathcal{V}$={there, bad, dull, moment, good, boring, awesome, actors, classic, story, fights, …} of size V=|$\mathcal{V}$|, we will represent a word with a one-hot vector in $\mathbb{R}^V$

The vector for “bad” = (0, 1, 0, …, 0, 0, 0)

The vector for “good” = (0, 0, 0, 0, 1, …, 0)

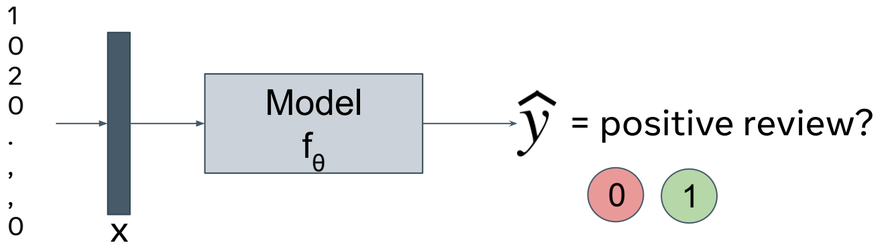
All except one position are zeros

Bag-of-words representation:

There is never a dull moment in this movie. Wonderful visuals and good actors.

Frequency of each word in the vocabulary

The vector of the sentence = (1, 0, 2, 0, …, 0)



Text representation

Bag-of-words representation:

Frequency of each word in the vocabulary

The vector of the sentence = (1, 0, 2, 0, …, 0) $\in \mathbb{R}^V$

The issues:

1. **Large vocabularies** mean large sparse vectors. (V≫10^5)
2. **Loss of word order** information.

   vector(“I like tagine but hate couscous”) = vector(“I hate tagine but like couscous”)

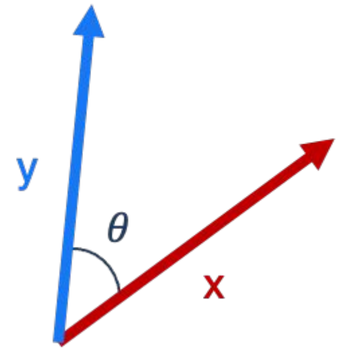
3. There is **no notion of similarity**:

   dot(W(‘bad’), W(‘boring’)) = dot(W(‘bad’), W(‘awesome’)) = 0,

   where dot(x, y) = $x^T y$ = ||x|| ||y|| cos(θ) is the dot product of vectors x & y

   We want our vectors to capture **semantic information** i.e.
   if it’s the same meaning it should be the same vector.

$x^T y$ is high if x & y point to the same direction



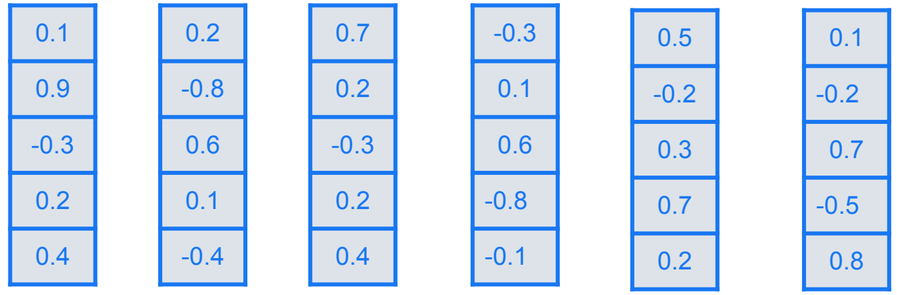
Text representation

The issues:

1. **Large vocabularies** mean large sparse vectors.

   (V≫10^5)

   ➢ We will use **dense** embeddings in $\mathbb{R}^d$ (d<<V).



2. **No notion of similarity**.

   We want to capture **semantic information**.

The Distributional Hypothesis:
Words that occur in the same contexts tend to have similar meanings

(Harris, 1954)

Solving 1+2 gave us **contextualized word vectors (or contextual embedding)**

How: The skip-gram model [Mikolov et al., 2013]

Text representation

Contextualized word vectors with skip-grams [Mikolov et al., 2013] (a simplified version for illustration purposes)

Since similar words appear in similar contexts, we will represent the word “**UM6P**” by these contexts from a training data.

Located in the “Mohammed VI Green City” in Benguerir, near Marrakech, **UM6P** applies a “learning by doing” approach
The project will leverage the expertise of INNOV’X, an innovation engine launched by **UM6P** in 2022 dedicated to building innovative and sustainable businesses and ecosystems
The 13th edition of the Roundtables of Arbois and the Mediterranean saw Morocco’s OCP Group and Mohammed VI Polytechnic University (**UM6P**) showcase progress on green hydrogen technologies, as well as the importance of such technologies for the institutions.

Morocco’s **UM6P** Bags Gold Medal at International Exhibition of Inventions in Geneva

UM6P’s Green Energy Park won an innovation award for its contributions to renewable energy research and development.

The UNITY team represented **UM6P** among ten schools from the African continent that participated in this international event.
This immersive visit at **UM6P** Campus is part of and “engagement course” for the students, to familiarize them with topics related to talent development, the needs for technology and innovation in the country,

The designation of **UM6P** by the members of the steering committee as the winner of the "Coup de Coeur" was motivated by the initiatives taken and carried out to develop professional equality in the workplace. **UM6P** was also congratulated and appreciated by the jury for its good practices.

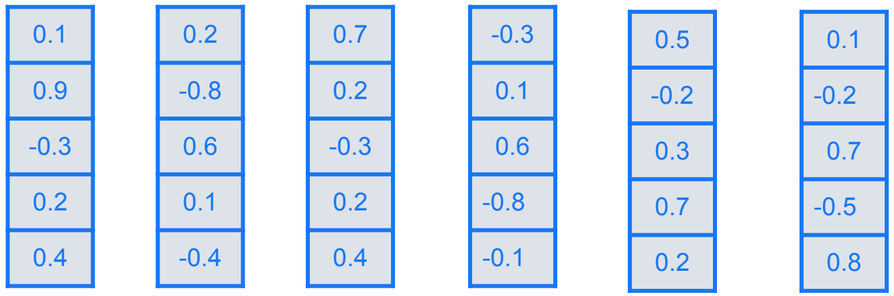
We will maximize the likelihood of observing all words surrounding the word UM6P

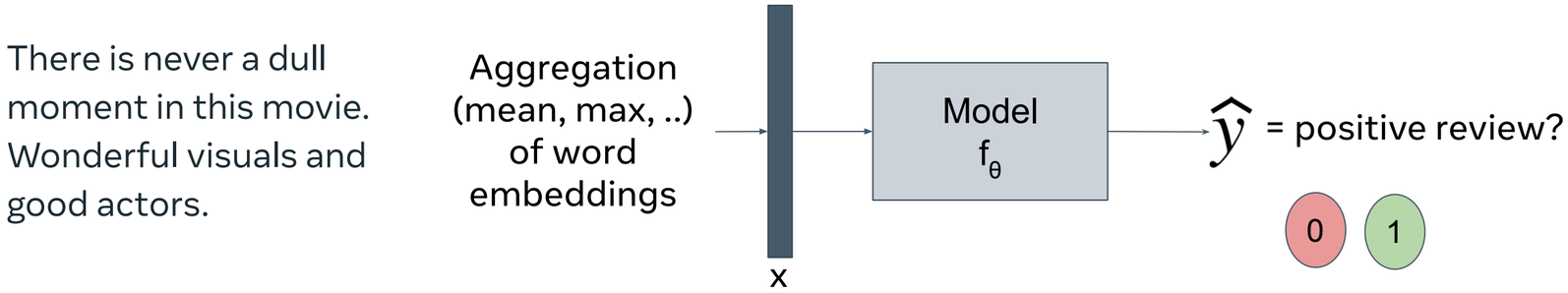
0.1	0.2	0.7	-0.3	0.5	0.1
0.9	-0.8	0.2	0.1	-0.2	-0.2
-0.3	0.6	-0.3	0.6	0.3	0.7
0.2	0.1	0.2	-0.8	0.7	-0.5
0.4	-0.4	0.4	-0.1	0.2	0.8

$$p(Benguerir|UM6P) = \frac{exp(vector(UM6P)^{T} vector(Benguerir))}{\sum_{word} exp(vector(UM6P)^{T} vector(word))}$$

Text representation

Contextualized word vectors with skip-grams [Mikolov et al., 2013] (a simplified version for illustration purposes)



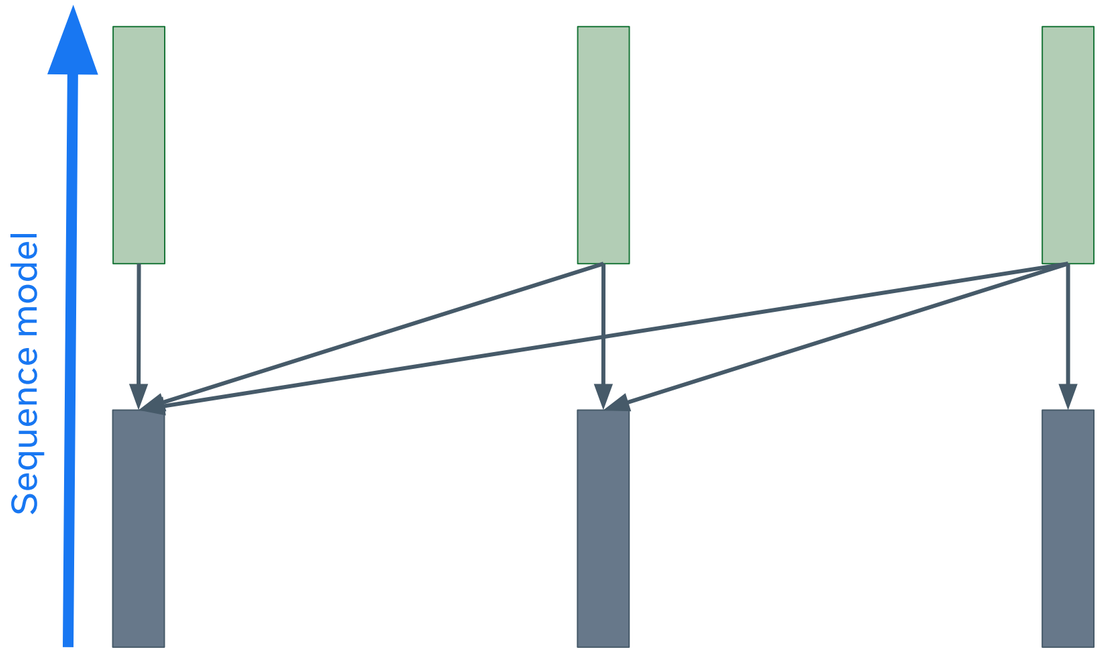


But we still haven't dealt with the **loss of word-order information**!

Text representation

Sequence models

Motivation: How to combine the words embeddings of a sentence (arbitrary length) into a meaningful vector representation?



context vectors

Eact vector represents the current word and all the previous words (context).

Some models include the future words as well.

Skip-gram vectors

How: Recurrent networks, 1D-convolutional models, attention models. For an introduction course check Lena Voita's "NLP Course for You" - https://lena-voita.github.io/nlp_course.html

Language models

Given a sequence x, the role of a language model is to estimate the joint probability p(x) i.e. to assess the plausibility or fluency of x.

With the chain rule we rewrite the joint probability as:

$$p(x_1, \dots, x_T) = \mathrm{p}(x_1).\,\mathrm{p}(x_2, |x_1) \dots\, \mathrm{p}(x_t, |x_1, \dots, x_{t-1}) \dots p(x_T|x_1, \dots, x_{T-1})$$

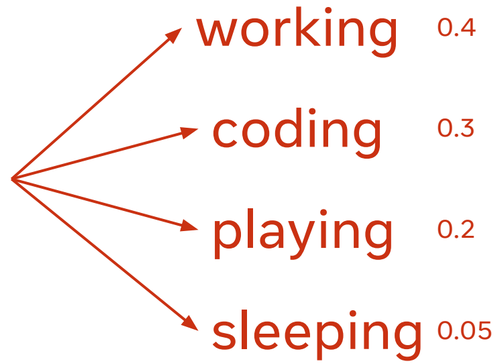
$$= p(x_1).\prod_{t=2}^{T} p(x_t|\, x_1, \dots, x_{t-1})$$

we compute the probability distribution of the next word x_t : $p(x_t \mid x_1, x_2 \dots, x_{t-1})$ and sample from it or pick the token with the highest probability (if greedy decoding).

Think of it as auto-complete

The hackathon participants spent the whole weekend __

x_1 x_2 x_3 x_4 x_5 x_6 x_7

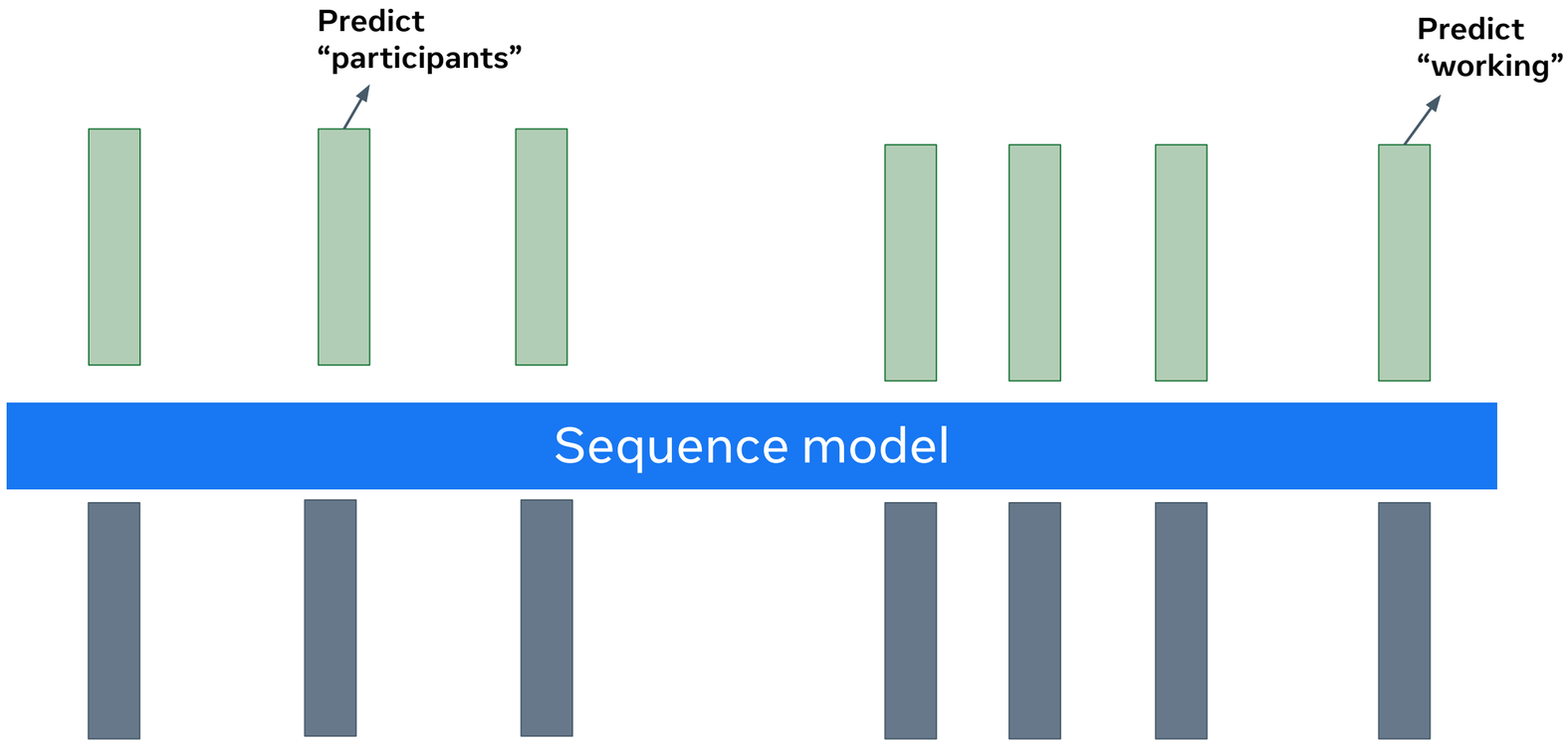


Language models

How are they related to sequence models?

We want to model $x_t \mid x_1\, x_2, \ldots, x_{t-1}$

If we have a **vector that summarizes $x_1\, x_2, \ldots, x_{t-1}$** then we can use it to predict what x_t could be. This vector is exactly what the output of a sequence model (encoder) is.



Conditional Language models

We want to model the **conditional joint probability of y | x**, where **y is a sequence** of tokens and **x is some conditioning context** (potentially a sequence itself).

Similar to language models, we decompose this probability with the chain rule:

$$p(y|x) = p(y_1|x).\prod_{t=2}^{T} p(y_t|\ y_1, \dots, y_{t-1}, x)$$

What is the probability of the next word, given the history of previously generated words AND conditioning context c.

Similar to LMs, except from the additional context usually processed with an encoder. This module + the LM decoder build what we call sequence-to-sequence models.

Conditional Language models

What NLP tasks do we use conditional LMs for?

The gas becomes thinner as you go farther from the Sun.

Source (**x**, <eng>)

Machine Translation (text-to-text)

كتزاد لكثافة ديال لغاز كلما بعدنا على شمس

Target (**y**, <ary>)

Speech Translation (speech-to-text or speech-to-speech)

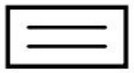
كتزاد لكثافة ديال لغاز كلما بعدنا على شمس

Target (**y**, <ary>)

Automatic speech recognition ASR

The gas becomes thinner as you go farther from the Sun.

Summarization



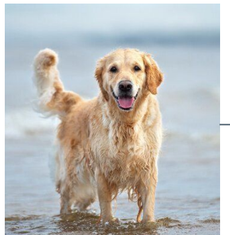


Image captioning

A dog standing in the ocean.

NLP:

It is the field of automatic (or semi-automatic) processing of human **languages**. It is focused on understanding human language at different **granularities** (characters, words, sentences, documents,…, etc.) and in different **modalities** (text, speech, visual, … etc.)

It’s not just LLMs!

Next: Some of our recent works in multilingual and multimodal NLP (speech+ text).

NLLB (No Language Left Behind)

North star goal: Develop a general-purpose **universal** machine translation model capable of translating between **any two** languages in various domains.

Google Translate supports **134** & Microsoft Translator supports **110, however,** there are more than **3000** written languages in the world. How can we break the 200 barrier?

Our work towards NLLB-200 was structured around 3 axes:

Data

Mining

Modeling

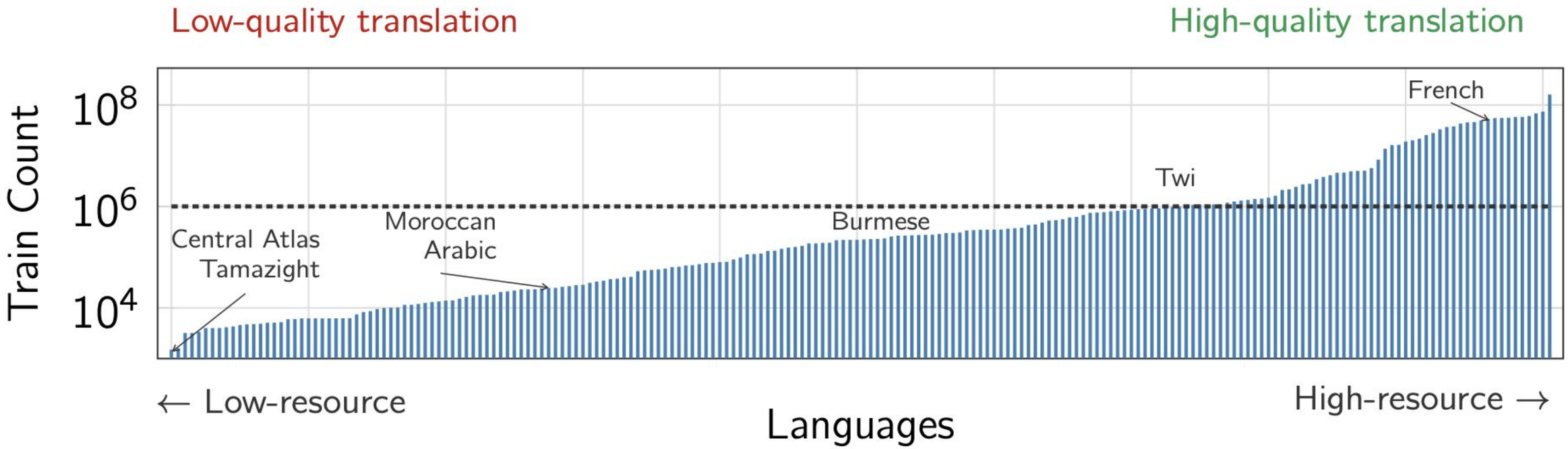
Mixture of Experts

Evaluation

NLLB (No Language Left Behind)

Problem: How can we collect enough training data for low-resource languages?

Bitexts data (pairs of source-target sentences) available to us per language



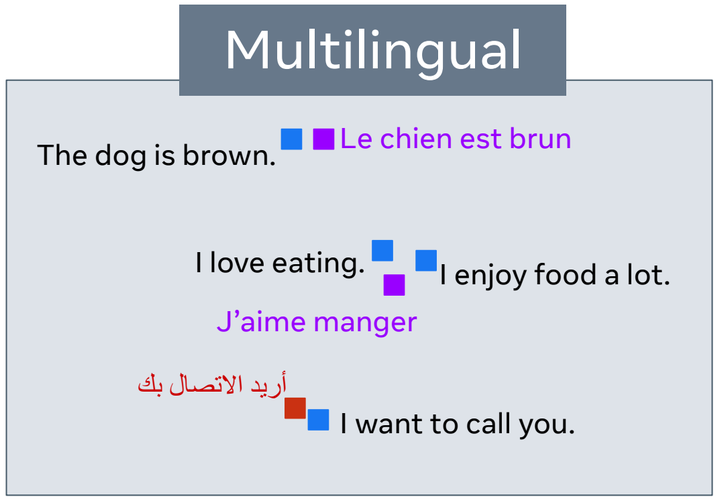
Two techniques to augment our data: (1) Back-translation, **(2) Bitext mining**.

NLLB (No Language Left Behind)

Bitext mining:

Multilingual Sentence Encoders to embed sentences and find semantically similar ones in different languages – see LASER (Artexte and Schwenk, 2019).

Sentences with similar meaning are ***close***.

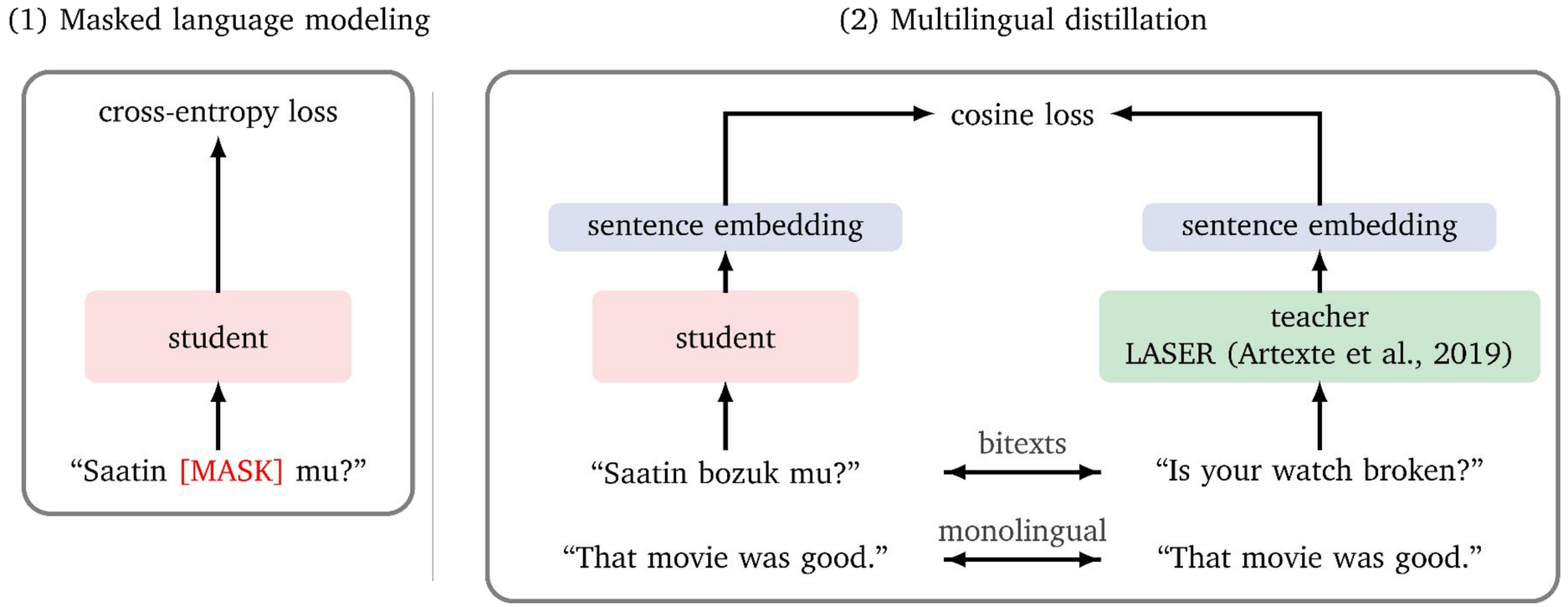


Sentences with similar meaning are ***close*** **independently of their language**

NLLB (No Language Left Behind)

Bitext mining:

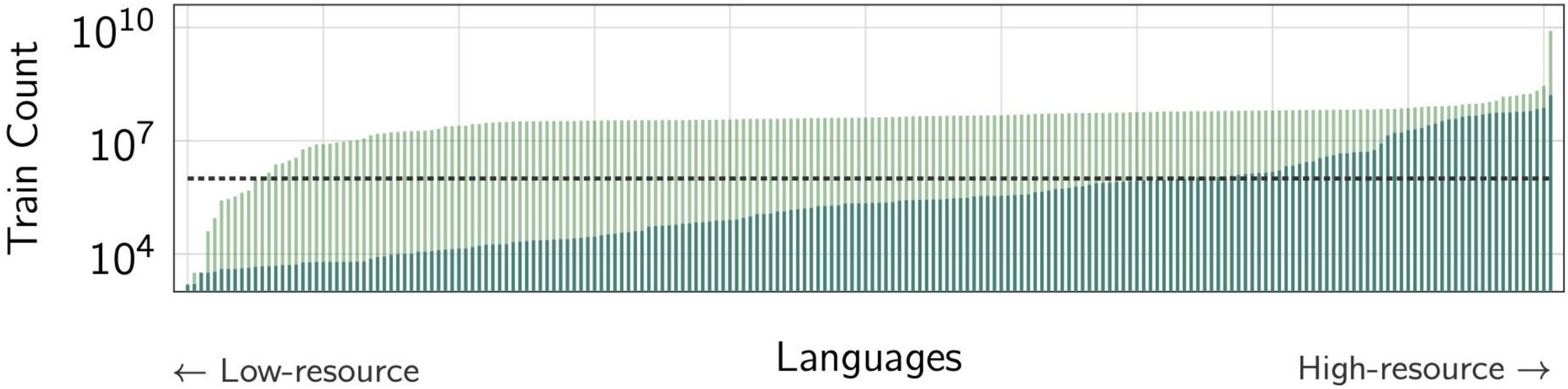
What if we have a new language and we want to encode it in the same multilingual space?



NLLB (No Language Left Behind)

Problem: How can we collect enough training data for low-resource languages?

With the addition of back-translated and mined data, most of the low-resource languages cross the threshold of 1M samples.

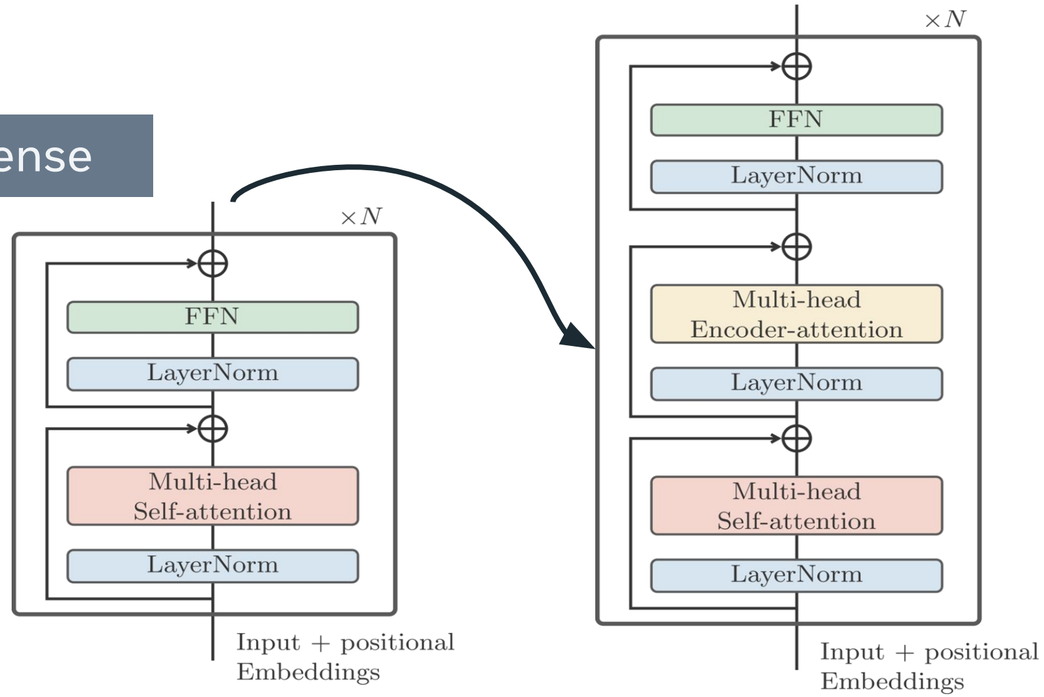


NLLB (No Language Left Behind)

The strength of Multinual MT is in leveraging **knowledge transfer** between languages. However, it also comes with **interference** between languages.

The MoE solution: A technique that allows for more parameters at an equivalent computational cost and for sparsely activated weights to be specialized in some languages.

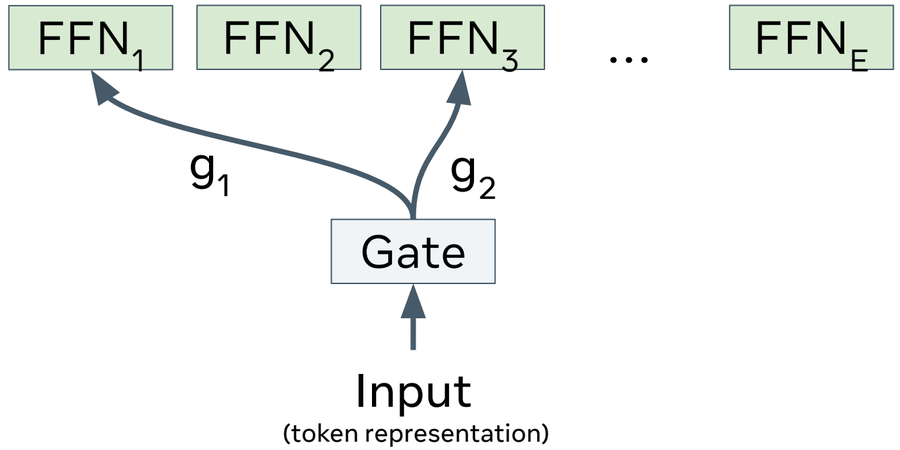
Dense



Encoder
Source sentence prefixed with <source_language>

Decoder
Target sentence prefixed with <target_language>

Sparsely Gated
Mixture of Experts (MoE)



Replace every other FFN in the Transformer model with an MoE FFN layer

NLLB (No Language Left Behind)

Results

Our final model significantly outperforms previous SOTA.

Flores-101

	eng_Latn-xx	xx-eng_Latn	xx-yy	Avg.
87 languages				
M2M-100	-/-	-/-	-/-	13.6/-
Deepnet	-/-	-/-	-/-	18.6/-
NLLB-200	**35.4**/52.1	**42.4**/62.1	**25.2**/43.2	**25.5**/43.5
101 languages				
DeltaLM	26.6/-	33.2/-	16.4/-	16.7/-
NLLB-200	**34.0**/50.6	**41.2**/60.9	**23.7**/41.4	**24.0**/41.7

+37%

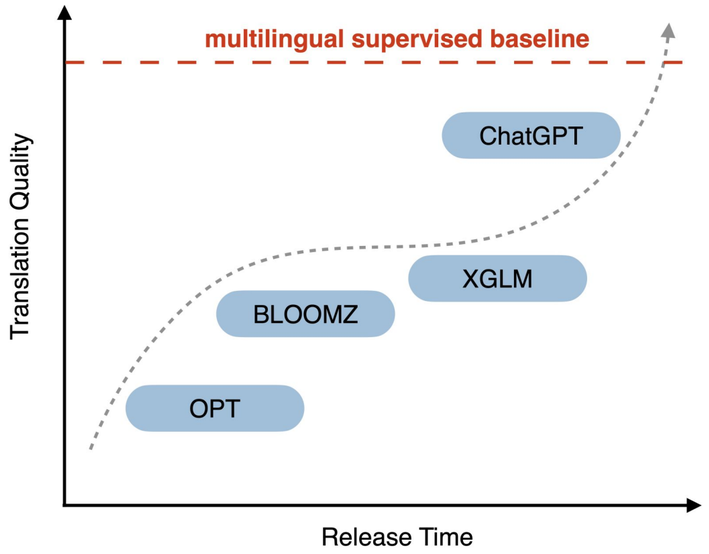
+44%

Flores-200

	eng_Latn-xx				xx-eng_Latn				xx-yy	Average
	all	high	low	v.low	all	high	low	v.low	all	all
chrF++	45.3	54.9	41.9	39.5	56.8	63.5	54.4	54.4	35.6	35.7
spBLEU	27.1	38.3	23.1	21.3	38.0	44.7	35.5	35.6	17.3	17.5

NLLB (No Language Left Behind)

Will LLMs replace supervised MMT models?



Recent studies (wip) have shown that **knowledge transfers poorly across languages in LLMs**: being correct on a specific question in English does not necessarily imply the LLM will also be correct on the same question in other languages.

[Submitted on 10 Apr 2023 (v1), last revised 2 May 2023 (this version, v2)]

Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, Lei Li

Download PDF

Large language models (LLMs) have demonstrated remarkable potential in handling multilingual machine translation (MMT). In this paper, we systematically investigate the advantages and challenges of LLMs for MMT by answering two questions: 1) How well do LLMs perform in translating a massive number of languages? 2) Which factors affect LLMs' performance in translation? We evaluate popular LLMs, including XGLM, OPT, BLOOMZ, and ChatGPT, on 102 languages. Our empirical results show that even the best model ChatGPT still lags behind the supervised baseline NLLB in 83.33% of translation directions. Through further analysis, we discover that LLMs exhibit new working patterns when used for MMT. First, prompt semantics can surprisingly be ignored when given in-context exemplars, where LLMs still show strong performance even with unreasonable prompts. Second, cross-lingual exemplars can provide better task instruction for low-resource translation than exemplars in the same language pairs. Third, we observe the overestimated performance of BLOOMZ on dataset Flores-101, indicating the potential risk when using public datasets for evaluation.

NLLB (No Language Left Behind)

Open-source!

- Project webpage: https://ai.facebook.com/research/no-language-left-behind/
- The Paper: https://arxiv.org/abs/2207.04672

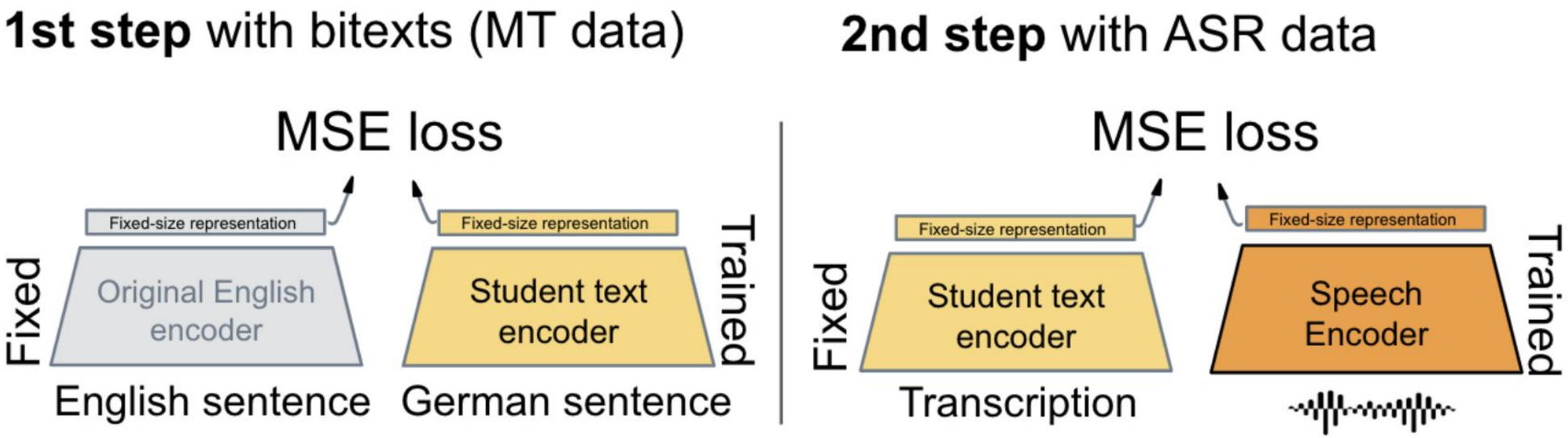
Codebases
- Modeling: https://github.com/facebookresearch/fairseq/tree/nllb
- LASER3: https://github.com/facebookresearch/LASER/blob/main/nllb
- Stopes (data and mining pipelines): https://github.com/facebookresearch/stopes/

Models
- Final NMT models: https://github.com/facebookresearch/fairseq/tree/nllb#multilingual-translation-models
- LASER3 encoders: https://github.com/facebookresearch/LASER/blob/main/nllb

Data
- Flores-200, NLLB-Seed, NLLB-MD, Toxicity-200: https://github.com/facebookresearch/flores
- Mined bitexts: https://huggingface.co/datasets/allenai/nllb

Ongoing work on multimodality (speech + text)

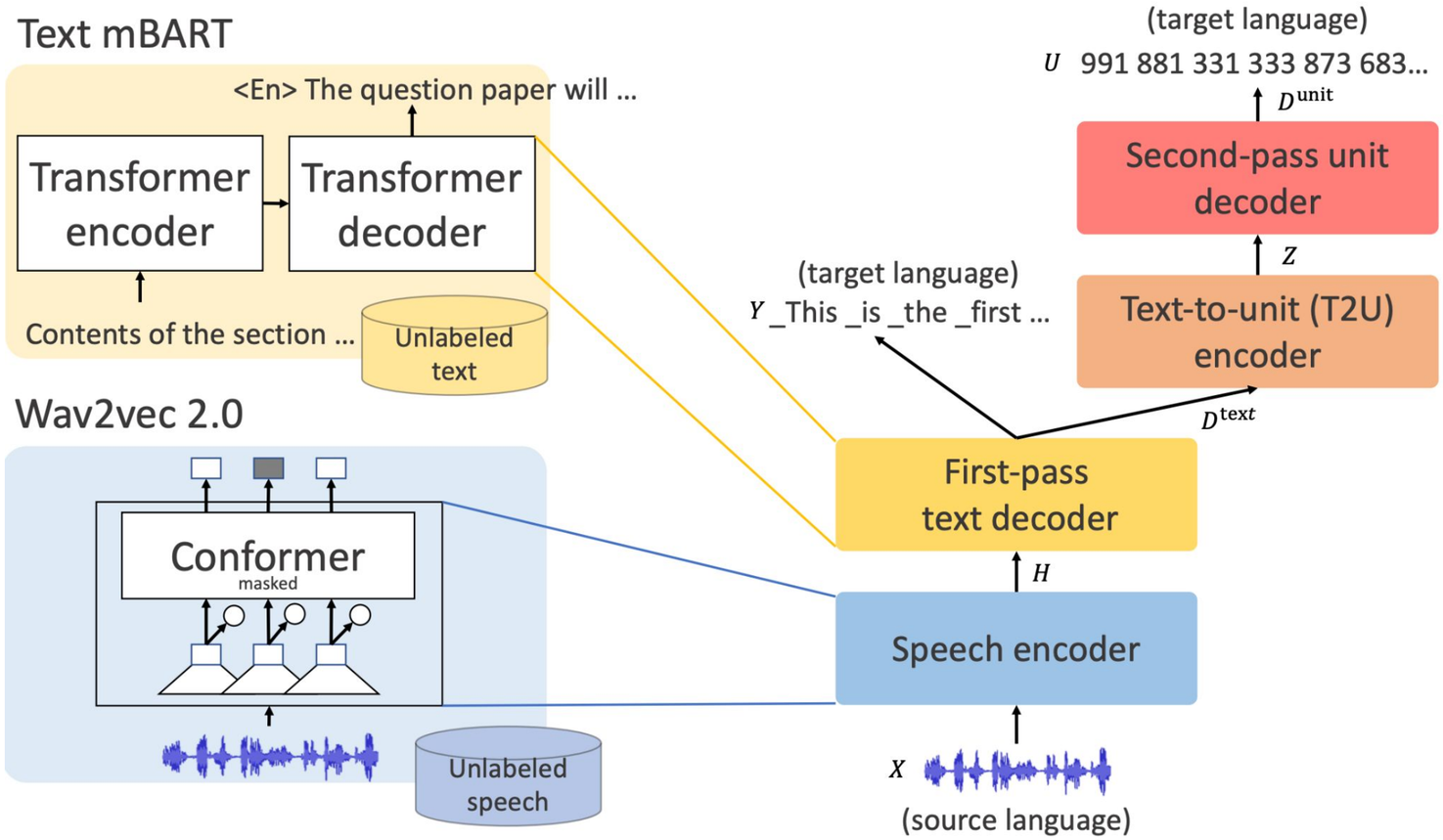
We extended LASER sentence embeddings to the speech modality with **SpeechMatrix** and **T-modules** (Duquenne et al. 2022).



Ongoing work on multimodality (speech + text)

Training **end-to-end multimodal models** (MT, ASR, S2T, S2ST) that are multilingual on both source and target sides (e.g. Whisper S2T does only translate into English)

We can already generate text and speech (units) with a two-pass decoder in UnitY (Inaguma et al. 2022)



Conclusion

There is more to NLP than LLMs.

The underlying modeling basics for these different tasks is the same.
So know your basics! The algorithmical basics.

There are a multitude of solutions to the same problem. As a researcher/engineer you hypothesise, then you use data to prove (or disprove) your hypothesis. And you iterate!