From Natural Language Processing to Transformer

Alexandre Allauzen and Florian Le Bronnec

01/10/2024

MILES
Machine Intelligence and Learning Systems

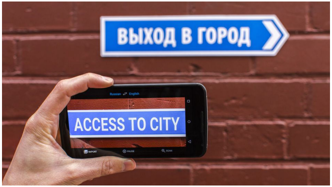
Dauphine UNIVERSITÉ PARIS | PSL

ESPCI PARIS | PSL

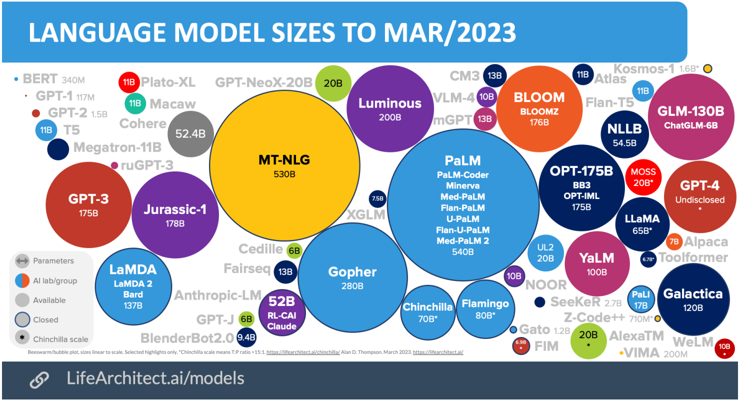
Roadmap

Outline

“Successful” applications of NLP

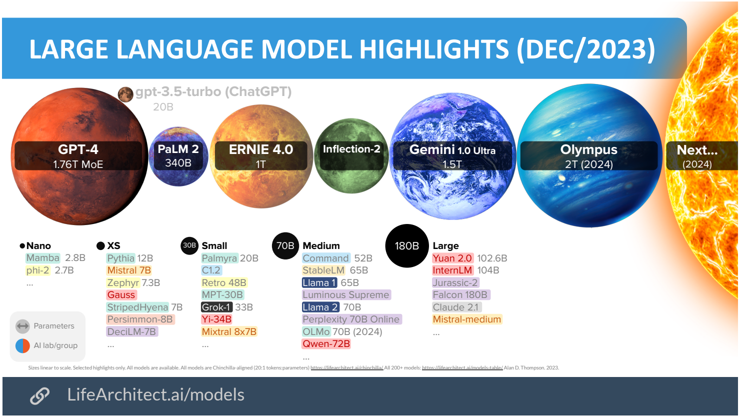


NLP or LLM ?



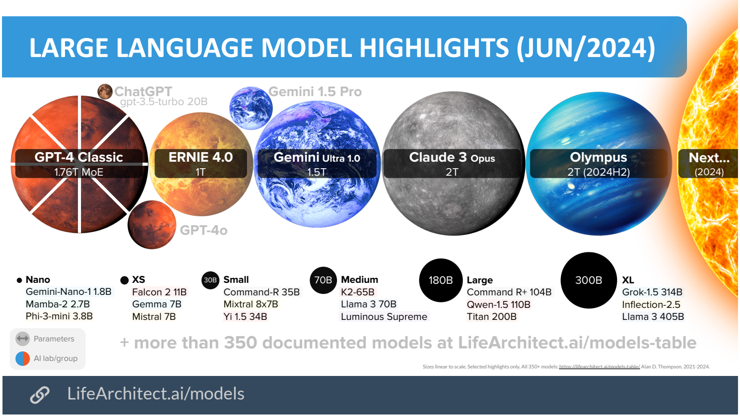
LANGUAGE MODEL SIZES TO MAR/2023
BERT 340M
GPT-1 117M
GPT-2 1.5B
11B T5
Megatron-11B
11B Plato-XL
11B Macaw
Cohere 52.4B
ruGPT-3
GPT-NeoX-20B 20B
Luminous 200B
CM3 13B
VLM-4 10B
mGPT 13B
BLOOM BLOOMZ 176B
11B Atlas
Kosmos-1 1.6B*
11B Flan-T5
GLM-130B ChatGLM-6B
NLLB 54.5B
GPT-3 175B
Jurassic-1 178B
MT-NLG 530B
7.5B XGLM
PaLM PaLM-Coder Minerva Med-PaLM Flan-PaLM U-PaLM Flan-U-PaLM Med-PaLM 2 540B
OPT-175B BB3 OPT-IML 175B
MOSS 20B*
GPT-4 Undisclosed *
LLaMA 65B*
7B Alpaca
6.7B* Toolformer
Parameters
AI lab/group
Available
Closed
Chinchilla scale
LaMDA LaMDA 2 Bard 137B
Cedille 6B
Fairseq 13B
Anthropic-LM
52B RL-CAI Claude
GPT-J 6B
BlenderBot2.0 9.4B
Gopher 280B
Chinchilla 70B*
Flamingo 80B*
10B NOOR
UL2 20B
YaLM 100B
SeeKeR 2.7B
PaLI 17B
Galactica 120B
Z-Code++ 710M*
Gato 1.2B
6.9B FIM
20B *
AlexaTM
WeLM 10B
VIMA 200M
Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/
LifeArchitect.ai/models

LLM or LLM ?



LARGE LANGUAGE MODEL HIGHLIGHTS (DEC/2023)
gpt-3.5-turbo (ChatGPT)
20B
GPT-4
1.76T MoE
PaLM 2
340B
ERNIE 4.0
1T
Inflection-2
Gemini 1.0 Ultra
1.5T
Olympus
2T (2024)
Next...
(2024)
Nano
Mamba 2.8B
phi-2 2.7B
...
XS
Pythia 12B
Mistral 7B
Zephyr 7.3B
Gauss
StripedHyena 7B
Persimmon-8B
DeciLM-7B
...
30B
Small
Palmyra 20B
C1.2
Retro 48B
MPT-30B
Grok-1 33B
Yi-34B
Mixtral 8x7B
...
70B
Medium
Command 52B
StableLM 65B
Llama 1 65B
Luminous Supreme
Llama 2 70B
Perplexity 70B Online
OLMo 70B (2024)
Qwen-72B
...
180B
Large
Yuan 2.0 102.6B
InternLM 104B
Jurassic-2
Falcon 180B
Claude 2.1
Mistral-medium
...
Parameters
AI lab/group
Sizes linear to scale. Selected highlights only. All models are available. All models are Chinchilla-aligned (20:1 tokens:parameters) https://lifearchitect.ai/chinchilla/ All 200+ models: https://lifearchitect.ai/models-table/ Alan D. Thompson. 2023.
LifeArchitect.ai/models

Still growing . . .



A great success

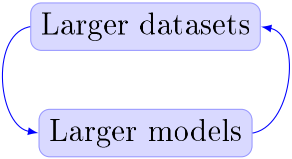
The reason of this success :

- availability of (very) large amount of training data
- progress of machine learning for NLP
- increase of computational power

$\rightarrow$ analyse and generate sentences, discourse, document

And a couple of decades

of diverse and intensive research



Outline

Language and languages

The "French" and strange distinction:

- "language" = a universal and innate capacity of humans to communicate and interact.
- "langue" = a "local" tool and you need to learn it

For more details see *Cours de linguistique générale* of Ferdinand de Saussure, 1916.

Diverse modalities

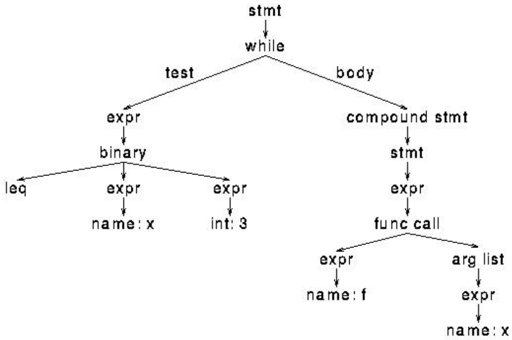
- Spoken and written forms
- Sign language
- Whistled language

Formal language

A Formal language is by construction explicit and non-ambiguous

- In the beginning was the grammar, like in programming language
- The grammar produces a unique parse for each statement, along with efficient (deterministic) parsing

```
while x <= 3:
    f(x)
```



Natural (human) language

The goal in the production and comprehension of natural language is communication.

For the speaker:

- Intention: what is the message ?
- Generation: translate the information into a sequence
- Synthesis: output the sequence (speech, signs, ...)

For the hearer:

- Perception
- Analysis
- Incorporation

Analysis: Syntax, Semantic, Pragmatic

Syntax concerns the proper ordering of words and its affect on meaning.

- the dog bit the boy.
- the boy bit the dog.
- bit boy dog the the.

Semantics concerns the (literal) meaning of words, phrases, and sentences.

- *plant*: an organism or a factory
- *plant*: the act of sowing

Pragmatics concerns the overall communicative and social context and its effect on interpretation.

- *The ham sandwich wants another beer.* (co-reference, anaphora)
- *John thinks vanilla.* (ellipsis)

Why is natural language implicit and ambiguous ?

> *Having a unique linguistic expression for every possible conceptualization would make language overly complex and linguistic expressions unnecessarily long.*
> *(R. Mooney)*

- Implicit enables conciseness (ellipsis, anaphora, ...), but entails potential ambiguities
- Unlimited expressivity requires flexible interpretation rules, the surface forms is just a part of the message.

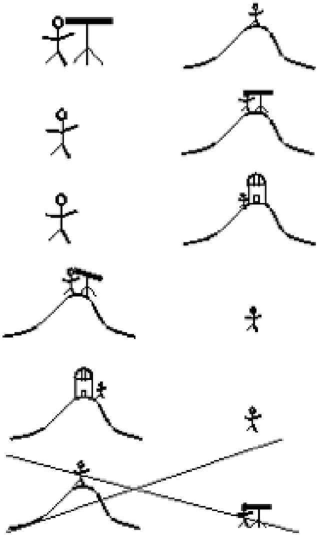
Cooking recipies

Remove the	stones	from the	cherries	and put them in the pie.
	sardines		cans	

Ambiguity

I saw the man on the hill with a telescope

(From the course of R. Mooney)



Ambiguity everywhere

Speech Recognition

"recognize speech" vs. "wreck a nice beach"

Syntactic Analysis

"I ate spaghetti with chopsticks" / "I ate spaghetti with meatballs"

Semantic Analysis

"The dog is in the pen"/ "The ink is in the pen"

Pragmatic Analysis (*The Pink Panther Strikes Again*)

Clouseau:	Does your dog bite?
Hotel Clerk:	No.
Clouseau:	(bowing down to pet the dog) Nice doggie. (Dog barks and bites Clouseau in the hand)
Clouseau:	I thought you said your dog did not bite!
Hotel Clerk:	That is not my dog.

Ambiguous, noisy and with great variability

Named entities and Idioms

- Where is A Bug's Life playing?
- Let It Be was recorded...
- Push the daisies

Non-canonical language

Great job @justinbieber! Were SOO PROUD of what youve done! U taught us 2 #neversaynever

Neologism

unfriend, retweet,Mary and Sue are sisters bromance, +1, ...

World knowledge

Mary and Sue are sisters / mothers

Headlines

- Hospitals are Sued by 7 Foot Doctors
- Kids Make Nutritious Snacks
- Iraqi Head Seeks Arms

World Wide

Beyond English, Mandarin, Arabic, ...

- Cultural heritage and digital humanities
- Language diversity

Outline

Morphology

Morphology: the field of linguistics that studies the internal structure of words.
A morpheme: the smallest linguistic unit that has semantic meaning (Wikipedia)

Examples

- independently: in + (depend + ent) + ly
- Mineralwasserflasche
- carried: carry + ed (the past tense)
- görüntülenebilir: görüntüle+n+ebil+ir
  visualize+passiv+can+be (can be visualized)

The morphological properties can carry a lot of information (gender, case, syntactic role, ...)

Part Of Speech (POS) Tagging

Sequence tagging

tag each word in a sentence with its part-of-speech (or morpho-syntactic property)

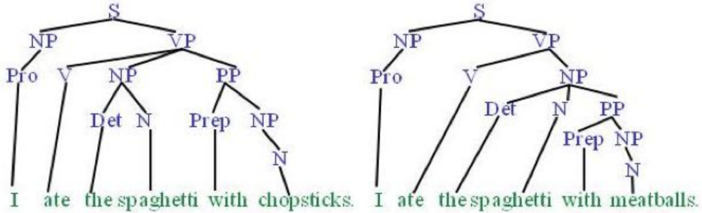
Sentence	POS-tags
er	PPER-case=nom gender=masc number=sg person=3
fürchtet	VVFIN-mood=ind number=sg person=3 tense=pres
noch	ADV
schlimmeres	NN-case=acc gender=neut number=sg
.	.

or in English:

I	ate	the	spaghetti	with	meatballs	.
Pro	V	Det	N	Prep	N	PUN

Syntactic parsing

Produce the correct syntactic parse tree for a sentence.



(R. Mooney)

Semantic analysis

Word Sense Disambiguation (WSD)

Infer the sense of each ambiguous word in a sentence

- Ellen has a strong interest in computational linguistics.
- Ellen pays a large amount of interest on her credit card.

Semantic Role Labeling (SRL)

Assign semantic role to words or phrases in a sentence (*e.g* agent, goal, or result)
(Agent Patient Source Destination Instrument)

John	drove	Mary	from	Austin	to	Dallas	in	his	DS Citroën
A		P		S		D			I

Important for many downtream tasks (Q&A, machine translation, …)

The 50's: two trends emerged

- Shannon explored probabilistic models of natural language (1951).
- Chomsky developed formal models of syntax, i.e. finite state and context-free grammars (1956).
- First computational parser developed at U-Penn as a cascade of finite-state transducers (Joshi, 1961; Harris, 1962).
- Bayesian methods developed for optical character recognition (OCR) (Bledsoe & Browning, 1959).

Who wrote ?

> *But it must be recognized that the notion of "probability of a sentence" is an entirely useless one, under any known interpretation of this term.*

The 60's

- Semantic network models of language for question answering (Simmons, 1965).
- First electronic corpus collected, Brown corpus, 1 million words (Kucera and Francis, 1967).
- Bayesian methods used to identify document authorship (The Federalist papers) (Mosteller & Wallace, 1964).

1964: The ALPAC report

The Automatic Language Processing Advisory Committee claimed that human translators are cheaper than the research in MT

The 70's: the Bayes rules and Markov models

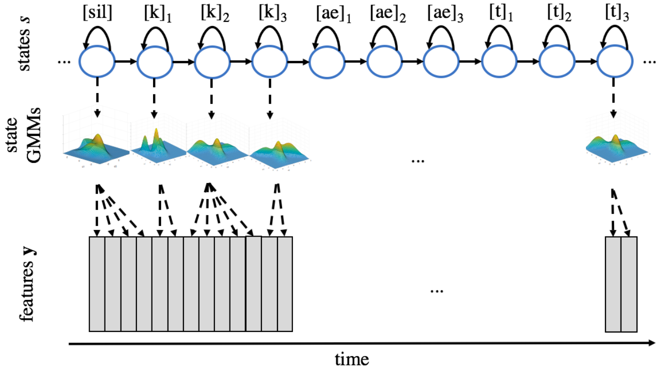
For speech recognition, machine translation, OCR, ...

$$P(\mathbf{W}|\mathbf{X}) = \frac{\overbrace{f(\mathbf{X}|\mathbf{W})}^{\text{acoustic}}\overbrace{P(\mathbf{W})}^{\text{prior}}}{P(\mathbf{X})}$$

- Acoustic likelihood: HMM and GMM
- Language model: prior model, originally a Markovian source

H.M.M rising

In the mid-70's: the spoken language as a Markovian source (Baker, 1975; Jelinek, 1976).



H.M.M and machine learning

A new mainstream

- For Syntactic analysis (POS tagging) in 1988
- Statistical Machine Translation (Brown et al., 1990)
- WordNet (Fellbaum & Miller, 90's)
- Other statistical models: SVM (Joachims,1999), CRF (Lafferty, 2001), Structured Perceptron (Collins, 2002).
- Unsupervised topic model, or Latent Dirichlet Allocation (Blei, 2003)

And in parallel, a new revolution

- Nettalk, a NNet "to read at loud"(Sejnowski et al., 1986)
- Then for POS tagging (Nakamura et al. 1986): the first embeddings !
- But really understood in (Bengio et al. 2001)

Outline

Language modelling task

A word prediction game

$$\begin{matrix} \text{time} & \text{goes} & \text{by} & \text{so} \\ w_1 & w_2 & \cdots & w_{i-1} \end{matrix} \longrightarrow w_i = ? \left\{ \begin{matrix} \text{a} \\ \vdots \\ \text{fastly} \\ \vdots \\ \text{slowly} \\ \vdots \end{matrix} \right.$$

A probability distribution over words

$$P(w_i | w_1^{i-1}),\ w_i \in \mathcal{V}$$

A probabilistic and generative model

$$P(w_1^L) = \prod_{i=1}^{L} P(w_i | w_1^{i-1}), \quad \forall i, w_i \in \mathcal{V}$$

Challenges

- Large vocabulary (from 10k to millions)
- Very sparse observation
- Large amount of available data but noisy, heterogenous, ...

A key task

Speech recognition and Machine Translation

Starting in the 80's [5, 1]:

[( / ), $\underbrace{w_1, \ \dots \ , w_{i-1}}_{\text{output prefix}}$] $\longrightarrow$ w_i ?

input context

And many others

- Handwritten character recognition, Text classification
- Chatbot, Question Answering

Count based model (from 80's to 2000)

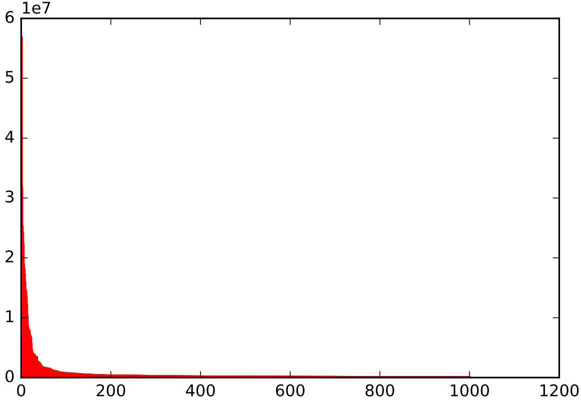
n-gram based model

$$P(w_i|w_1^{i-1}) \approx P(w_i|\underbrace{w_{i-n+1}^{i-1}}_{\substack{n-1 \\ \text{last words}}}) = \frac{c(w_i|w_1^{i-n+1})}{c(w_1^{i-n+1})}$$

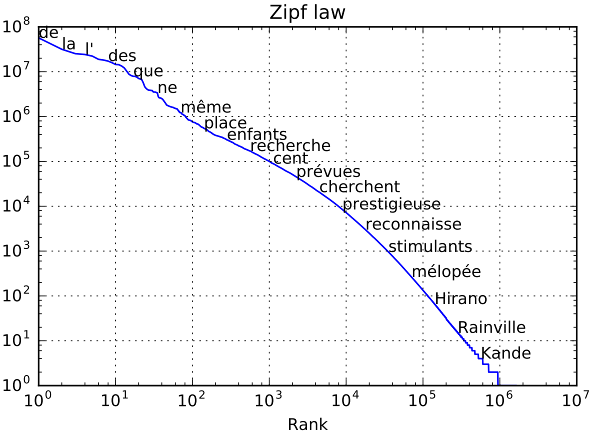
Lack of generalization

- smoothing methods as a workaround
- but no similarity between words

The Zipf law



The Zipf law - 2



A second life as an unsupervised learning task

"Language Models are Unsupervised Multitask Learners"

- Leveraging the huge amount of unlabeled texts
- To pre-train word representations
- Along with their contextualization at the sentence level
- That can be fine-tuned for downstream tasks

A profusion of architectures

- Starting with convolution networks [2]
- More recently ELMo and ULMFit with LSTM [6, 4]
- BERT and GPT with Transformers [3, 7]

Outline

Today: the main tasks

- Text classification

$$\mathbf{X} = (w_1, w_2, ..., w_L) = w_1^L \longrightarrow y \text{ the class}$$

- Word Tagging

$$\mathbf{X} = (w_1, w_2, ..., w_L) \longrightarrow (y_1, y_2, ..., y_L)$$

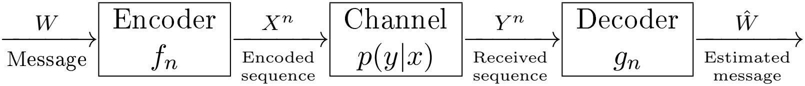
- Sentences (or relation) classification

$$(\mathbf{X}_1, \mathbf{X}_2) \longrightarrow y$$

- Conditionnal Generative Model

$$P(\mathbf{W}|\mathbf{X})$$

Encoder / Decoder



Outline

Outline

[1] Peter F. Brown et al. "A Statistical Approach to Machine Translation". In: *Computational Linguistics* 16.2 (1990), pp. 79–85. URL: citeseer.nj.nec.com/brown90statistical.html.

[2] Ronan Collobert and Jason Weston. "A unified architecture for natural language processing: deep neural networks with multitask learning". In: *Proceedings of the International Conference of Machine Learning (ICML)*. Helsinki, Finland: ACM, 2008, pp. 160–167. ISBN: 978-1-60558-205-4. DOI: http://doi.acm.org/10.1145/1390156.1390177.

[3] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[4] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 328–339. URL: http://aclweb.org/anthology/P18-1031.

[5] Frederick Jelinek. "Continuous Speech Recognition by Statistical Methods". In: *Proceedings of IEEE* 64.4 (1976), pp. 532–556.

[6] Matthew E. Peters et al. "Deep contextualized word representations". In: *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2018.

[7] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).