

From Natural Language Processing to Transformer

Alexandre Allauzen and Florian Le Bronnec

Winter 2024



Roadmap

Introduction

A bit of language, of ambiguity and diversity

Old Tasks and Milestones

Neural Language modelling

From language models to large language models

References

Outline

Introduction

A bit of language, of ambiguity and diversity

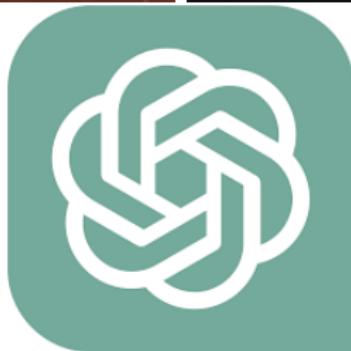
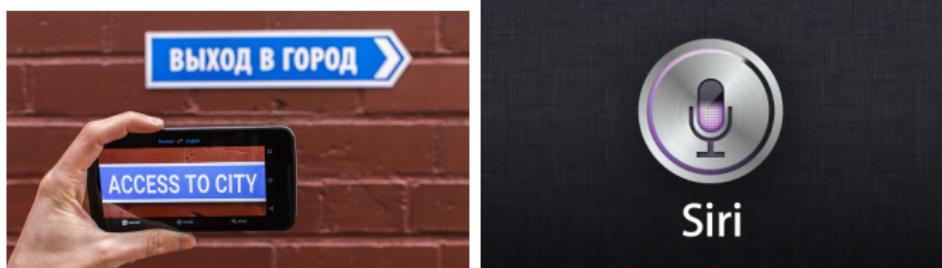
Old Tasks and Milestones

Neural Language modelling

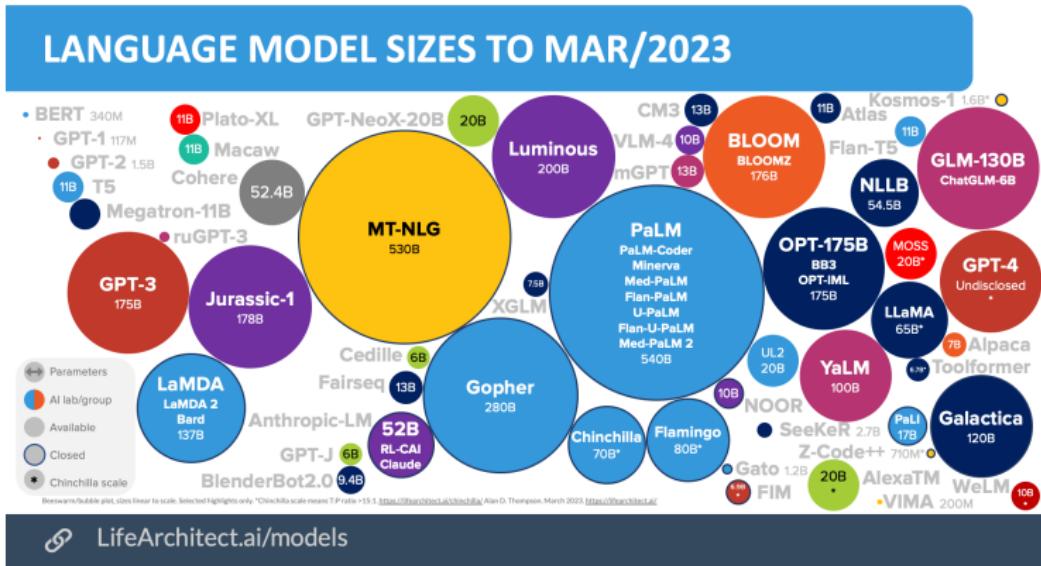
From language models to large language models

References

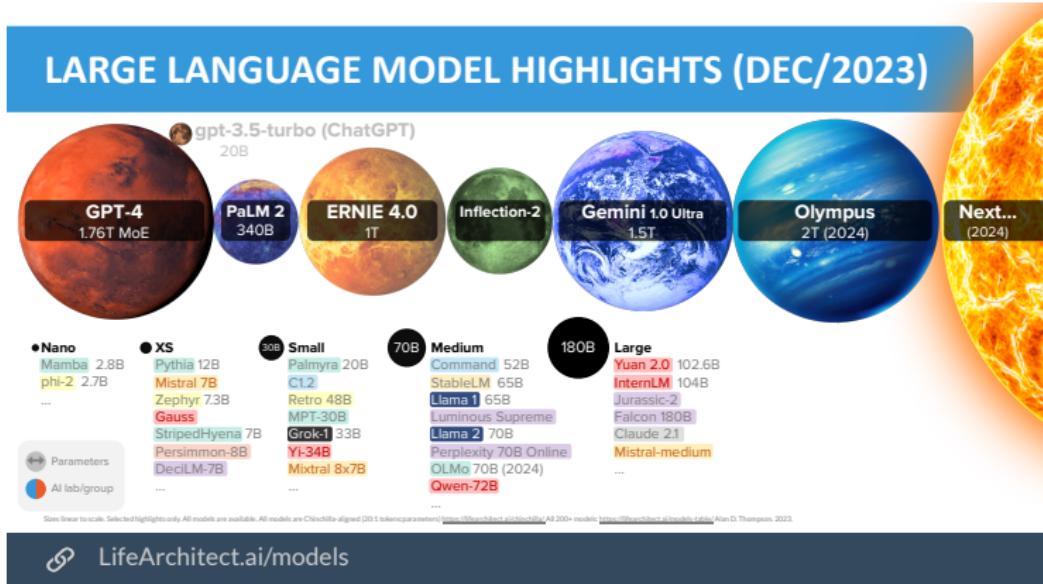
“Successful” applications of NLP



NLP or LLM ?



LLM or LLM ?



A great success

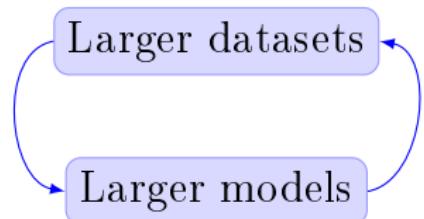
The reason of this success :

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

→ analyse and generate sentences, discourse, document

And a couple of decades

of diverse and intensive
research



Outline

Introduction

A bit of language, of ambiguity and diversity

Old Tasks and Milestones

Neural Language modelling

From language models to large language models

References

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.

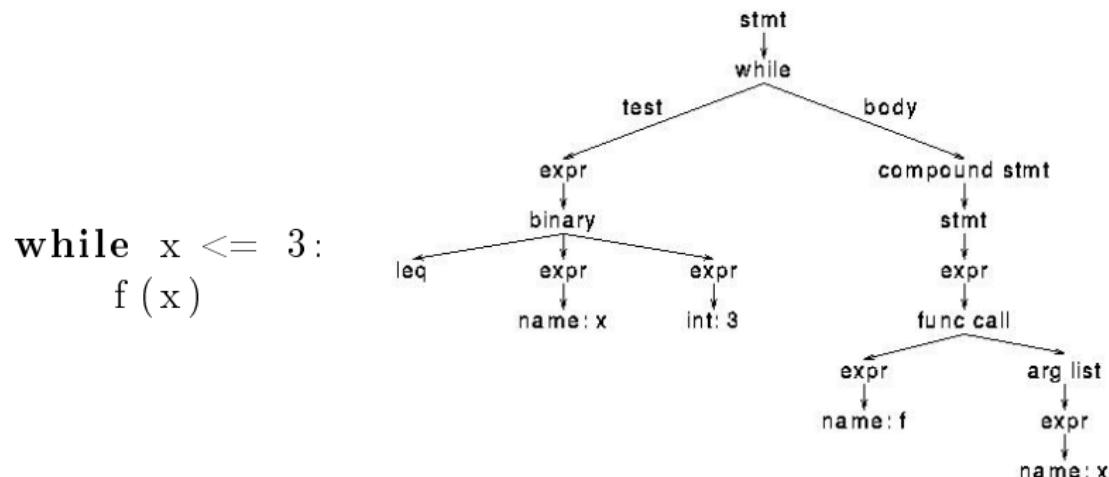
Diverse modalities

- 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



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 effect 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.

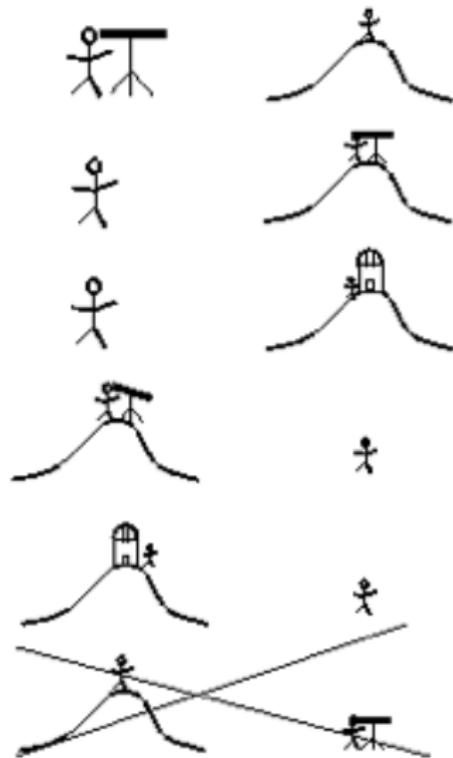
Cooking recipes

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

Introduction

A bit of language, of ambiguity and diversity

Old Tasks and Milestones

Neural Language modelling

From language models to large language models

References

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)

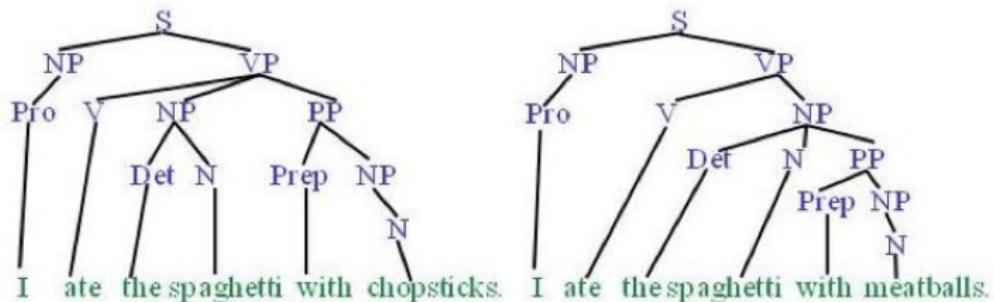
Sentence	POS-tags
er	PPER-case=nom gender=masc number=sg person=3
fürchtet	VVFIN-mood=ind number=sg person=3 tense=pres
noch	ADV
schlimmeres	NN-case=acc gender=neut number=sg
.	.

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 downstream 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

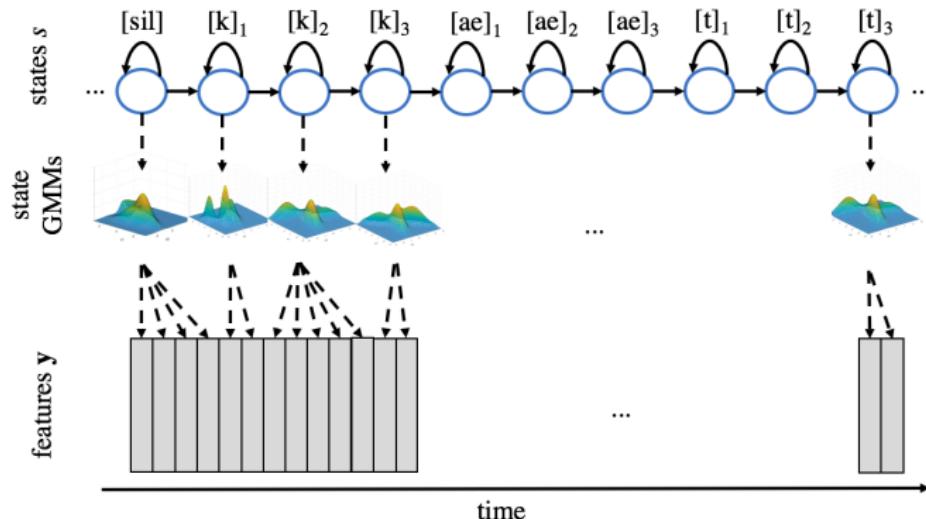
For speech recognition, machine translation, OCR, ...

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

- 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

Introduction

A bit of language, of ambiguity and diversity

Old Tasks and Milestones

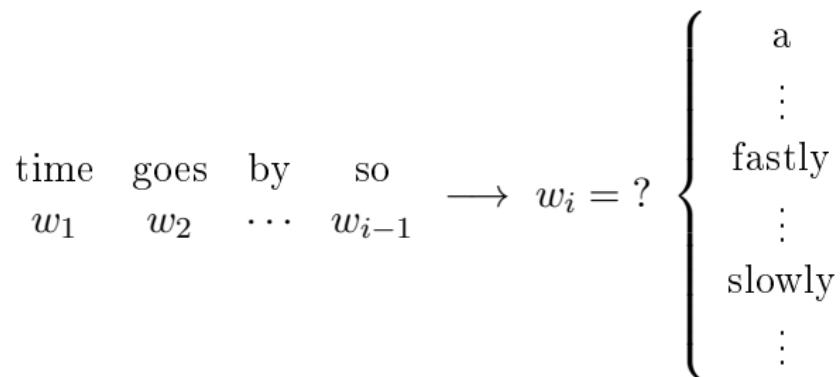
Neural Language modelling

From language models to large language models

References

Language modelling task

A word prediction game



A probability distribution over words

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

A probabilistic and generative model

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

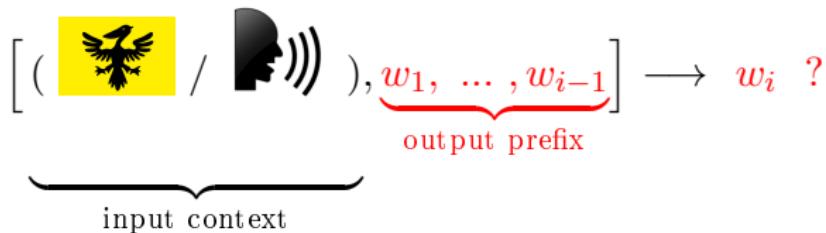
Challenges

- 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]:



And many others

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

Count based model (from 80's to 2000)

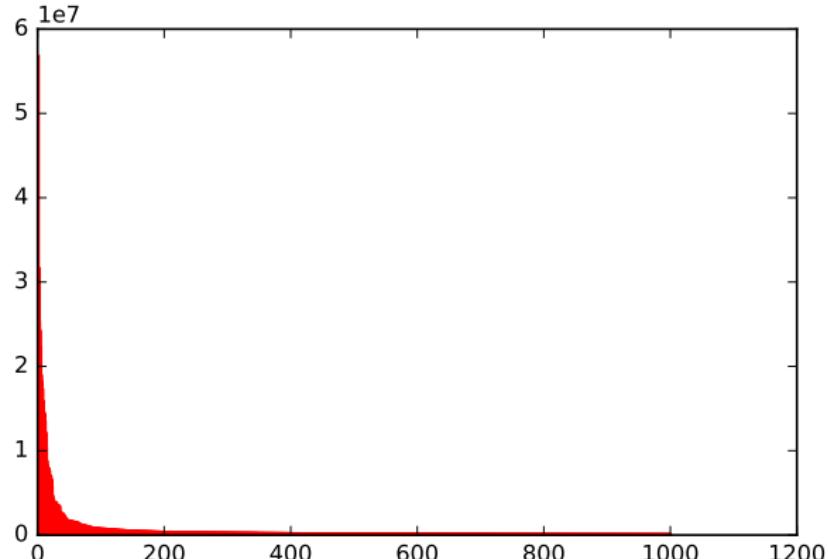
n-gram based model

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

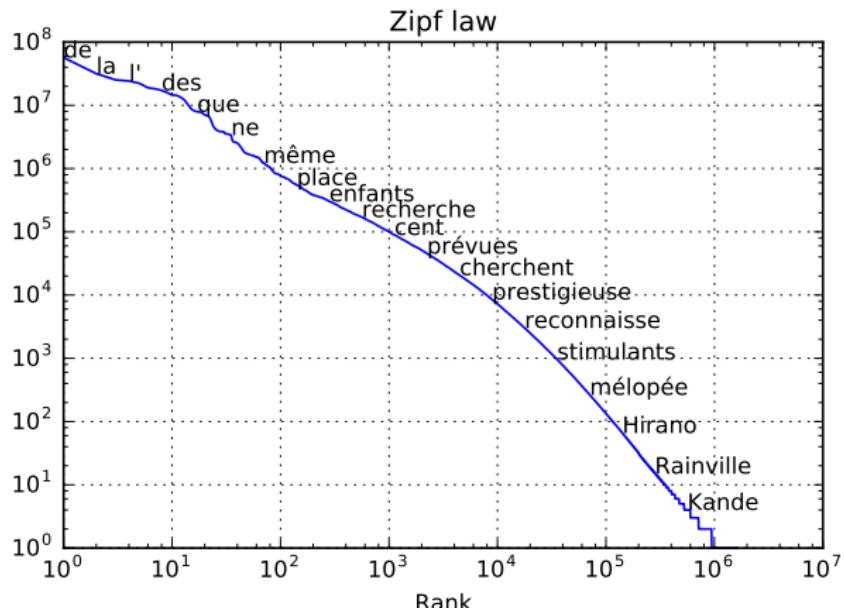
Lack of generalization

- 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

Introduction

A bit of language, of ambiguity and diversity

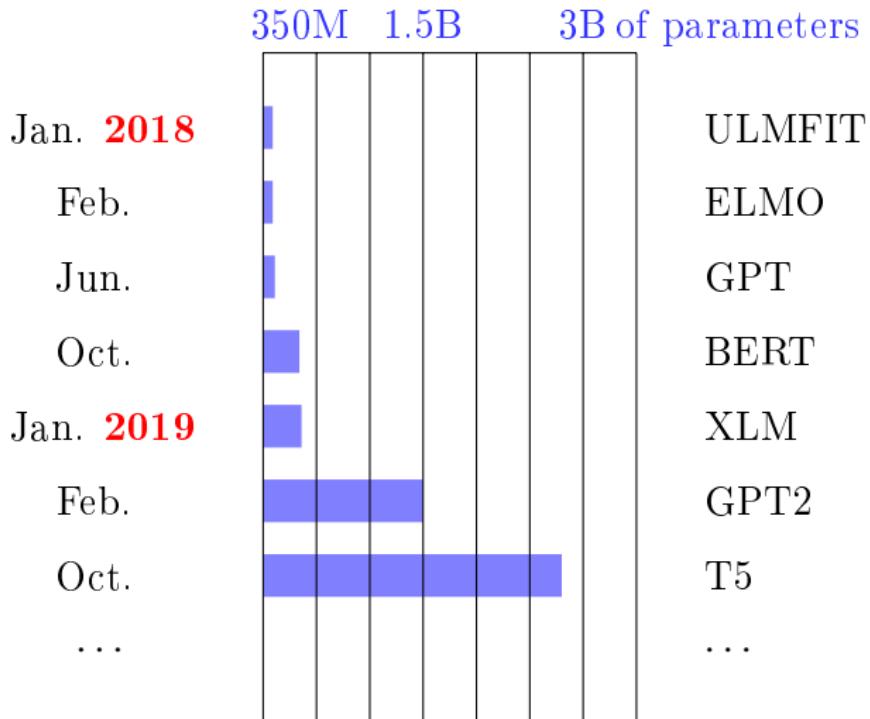
Old Tasks and Milestones

Neural Language modelling

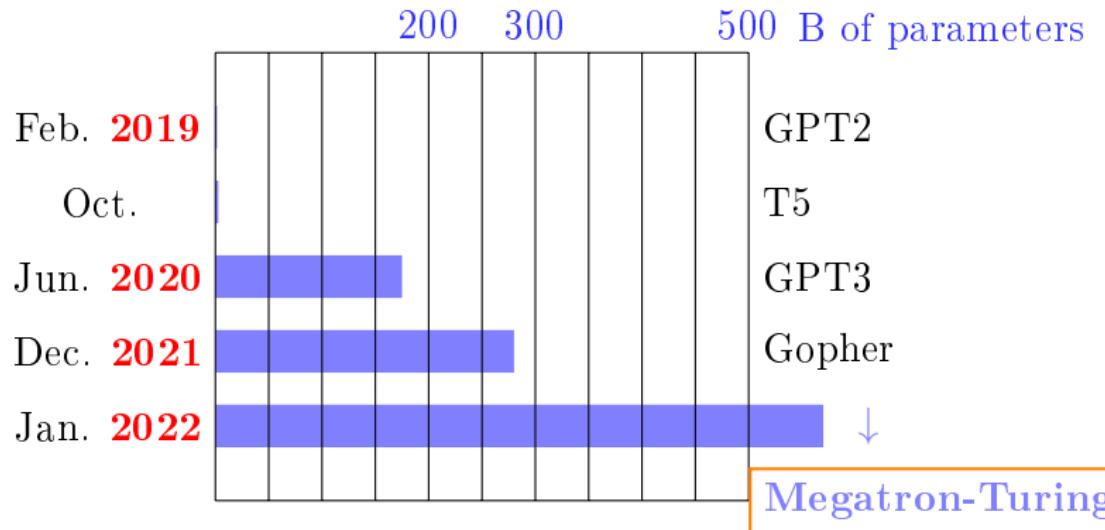
From language models to large language models

References

In a few dates



Bigger is ...

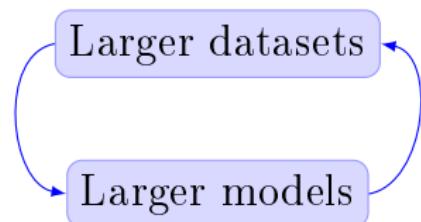


- GLaM (Dec. 2021) more than 1T of parameters
- And Llama, PALM, Mistral, Mixtral, ...

Why we need such resource ?

Variability in written languages

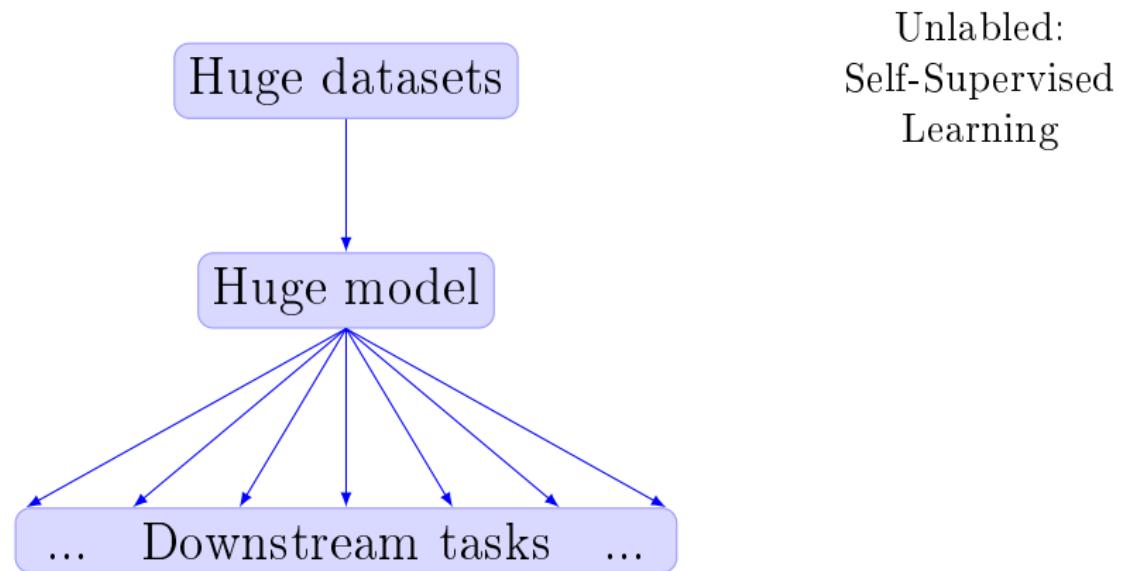
- Domain, style, topic, social context, the medium, ...
- For many language pairs
- For many languages, (all ?)
- Multi-modal (image captioning, data to text, SQL to text, ...)



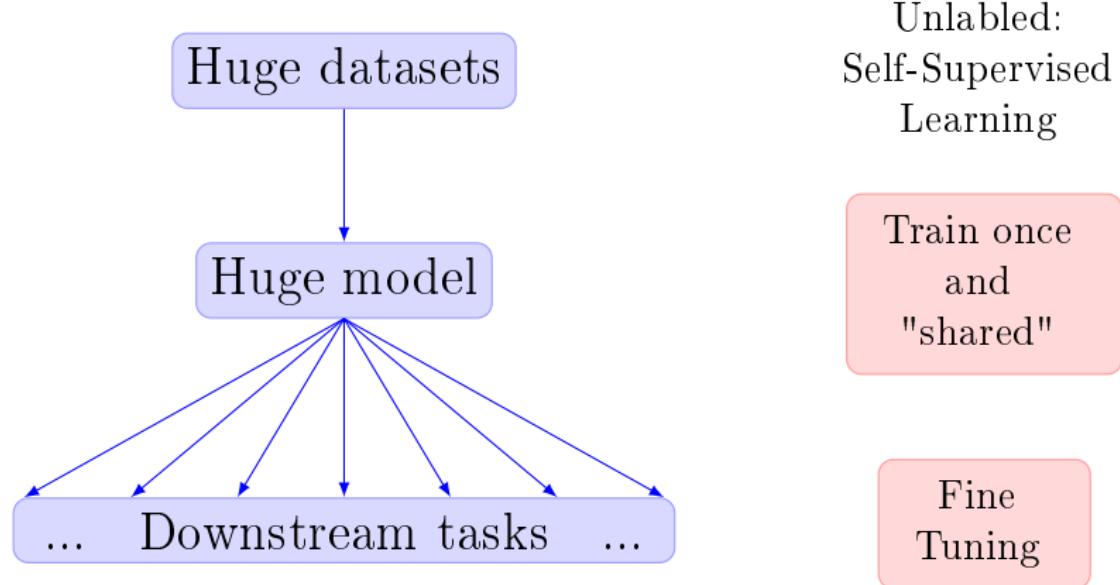
And it is not over !

What about speech processing ?

The promise of a new life cycle (resource-aware)



The promise of a new life cycle (resource-aware)



Beyond wishful thinking

What are the issues ?

- The amount of resources (CPUs, GPUs, storage, bandwidth)
- The experimental cost
- And the growth is unsustainable !

A new research perspective

- The design of resources aware architecture
- Versatility and performance
- The prior knowledge and better learning principles
- Seeking Invariance and symmetries
- ...

Today: the main tasks

- Text classification

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

- Word Tagging

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

- Sentences (or relation) classification

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

- Conditionnal Generative Model

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

Outline

Introduction

A bit of language, of ambiguity and diversity

Old Tasks and Milestones

Neural Language modelling

From language models to large language models

References

- [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).