



Bases de TAL

Master 1 Sciences du Langage
Parcours Industries de la Langue – Linguistic Data Sciences
Martina Barletta
3-6 septembre 2024

Licence CC BY-NC-SA

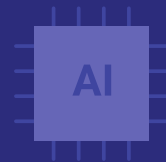
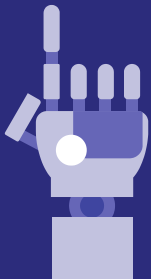




Table des contenus

01

Introduction

au Traitement
Automatique des Langues

02

Histoire du TAL

Du rapport ALPAC à
ChatGPT

03

Tâches

Et campagnes
d'évaluation

04

Evaluation

Comment on évalue des
systèmes TAL et
l'annotation des données ?

05


Métriques

Pourquoi et comment
évaluer correctement ?

06

Exposé

sur les thématiques
du cours



Doctorante au laboratoire LIDILEM
Annotation de la coréférence dans un corpus d'écrits
scolaires en français, italien et espagnol

- Linguistique des corpus
- Corpus écrits
- Coréférence
- Apprentissage des L1/L2



martina.barletta@uni-grenoble-alpes.fr

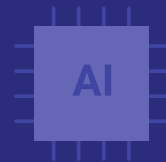
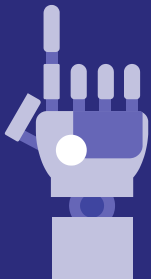
<https://martinabarletta.github.io/martina-barletta//teaching/> pour les diapos
(site en construction)



01

Introduction

Qu'est-ce que c'est le TAL ?



Industries de la langue ?

Scannez le QR code à l'écran ou connectez-vous avec le lien suivant au sondage : <https://www.menti.com/al13odojx7mn>

Code : **1762 2872**



Industries de la langue

Traducteurs automatiques, dictionnaires informatisés, correcteurs automatiques

Prédiction de mots

Moteurs de recherche

Recommender systems

Classification automatique des informations

Recherche d'informations

Analyse d'opinion ou de sentiments

Génération automatique de textes, résumé automatique

Synthèse vocale, reconnaissance de la parole

Chatbots (service vocaux, assistants vocaux interactifs)

Robotique

Produits pour la communication augmentée

...

À la base de ces produits, on retrouve
le **TRAITEMENT AUTOMATIQUE DES LANGUES**

Traitement Automatique de la Langue

- Élaboration de programmes informatiques manipulant la forme langagière et capables, à travers la forme, de traiter le sens associé, pour résoudre des problèmes spécifiques
- Le but du TAL est de faire en sorte que les problèmes mettant en jeu des informations langagières deviennent calculables
- Problème central du TAL : **Traiter l'ambiguïté**





**Le langage est
redondant, implicite
et ambigu**



Le langage est ambigu, implicite

Ambiguïté : un même segment linguistique peut se prêter à deux interprétations mutuellement exclusives (Kerbrat-Orecchioni, 2005)

Exemple	Type d'ambiguïté
C'est un <u>vol</u> très risqué	Sémantique → Homonymie
Je <u>loue</u> pour l'année un appartement à Grenoble	Sémantique → Polysémie
Couvent (nom ou verbe)	Catégorielle
La petite porte le voile	Syntaxique catégorielle
Je vois un homme sur la colline avec un télescope	Syntaxique structurale

L'implicite dans le langage

ARTS

Un Caravage a-t-il été découvert dans un grenier en France ?

Le tableau trouvé en 2014 dans la région de Toulouse est estimé 120 millions d'euros. Les spécialistes ont trente mois pour approfondir l'expertise.

Connaissance du monde
Métonymie

Éducation et enseignement supérieur

Les dossiers chauds d'une rentrée scolaire inédite

Métaphore

Amériques

La Maison Blanche, un vieux rêve pour la vice-présidente

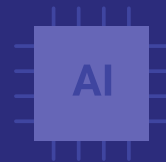
Connaissance du monde



02

Histoire du TAL

Du rapport ALPAC à ChatGPT



Traduction Automatique et Guerre Froide

1950 – USA – développement d'applications pour la traduction automatique

1954 – **Georgetown-IBM experiment**

Traduction automatique de 60 phrases russe → anglais

On préconisait la traduction automatique comme
problème déjà résolu dans l'espace de 3/5 ans...



“дух бодр, плоть же немощна”
“The spirit is willing, but the flesh is weak”

RU → EN → RU

“The vodka is strong, but the meat is rotten”
“водка хорошая, но мясо протухло”

(Exemple apocryphe non attesté)



“Out of mind, out of sight”

EN → RU → EN

“Invisible idiot”

cité par John Hutchins, Harper's Magazine, August 1962
"The whisky was invisible", or Persistent myths of MT



1966 - Rapport ALPAC (I)

ALPAC

Automatic Language Processing Advisory Committee

Objectifs

Evaluer les progrès des derniers dix ans en linguistique computationnelle et en particulier en traduction automatique

Résultats du rapport

La recherche en TA n'était pas suffisante, il faut recentrer les recherches sur la linguistique computationnelle de base d'abord → réduction des financements en TA

Premier HIVER DE L'IA

1966 - Rapport ALPAC (II)

- Pas la peine d'investir plus dans la TA (marché relativement petit, traductions automatiques encore fortement erronées)
- S'orienter dans le développement d'outils de support à la traduction (dictionnaires numériques)

Passage de la Traduction Automatique
à la Traduction Assisté par l'Ordinateur

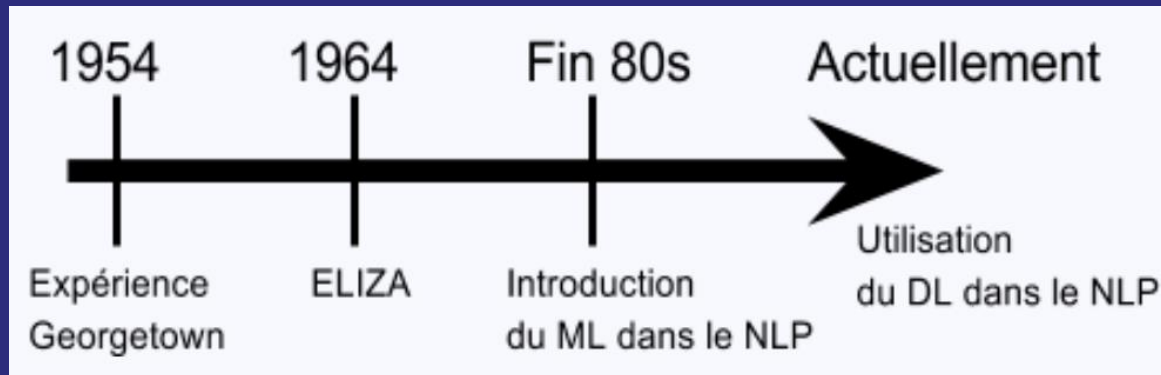
→ Naissance de la Linguistique Computationnelle
et de la Linguistique des Corpus

Périodisation

Période Symbolique (1950-1990)

Période Statistique (1990-2010)

Période Neuronale (2010-...)

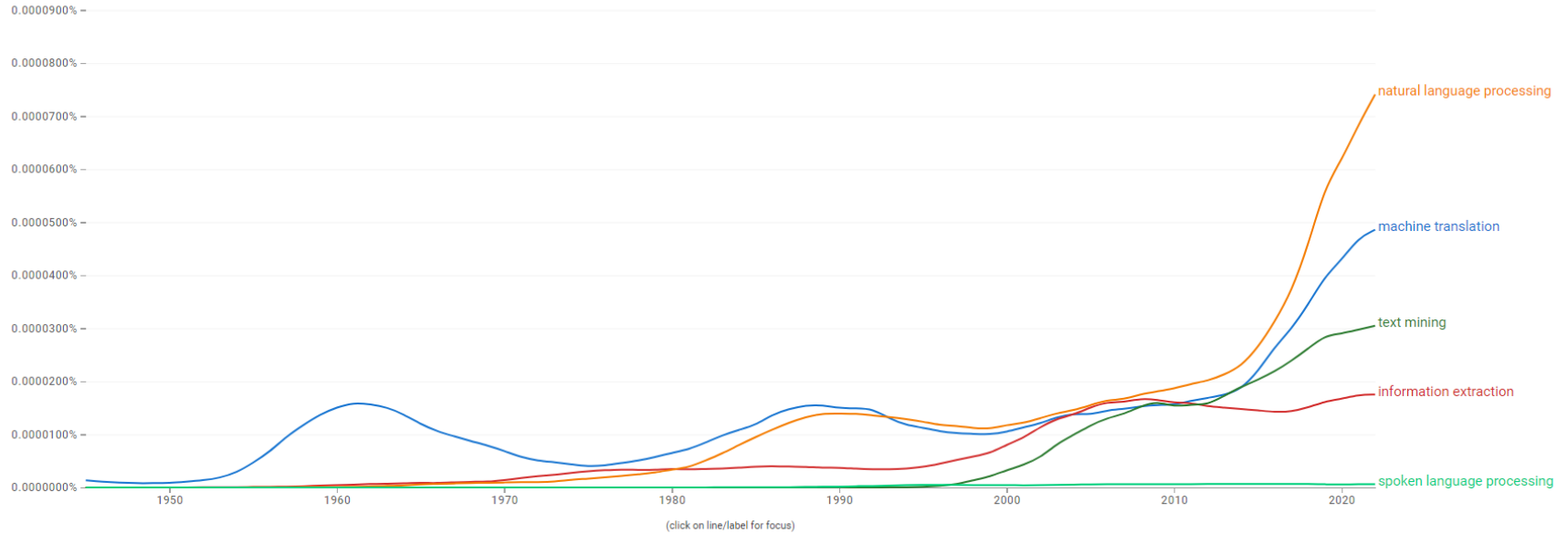


Périodisation

1. Early explorations (1940-1969)
2. Hand-build systems (1970-1992)
3. Statistical or probabilistic NLP (1993-2012)
4. Deep Learning or Artificial Neural Networks
Unsupervised or Selfsupervised
Reinforcement learning (2013-now)

Manning, C. D. (2022). Human Language Understanding & Reasoning. *Daedalus*, 151(2), 127-138. https://doi.org/10.1162/daed_a_01905

Google Books Ngram Viewer



Période symbolique

1950-1992 ca

Symbolique → les connaissances du système sont représentées sous forme de **règles**, décrite à travers des **symboles** lisibles par l'humain (langage formel – Chomsky)

Applications :

SHRDLU (Winograd, 1968-1970)

ELIZA (Weizenbaum, 1964-1967)

Période symbolique

Test de Turing

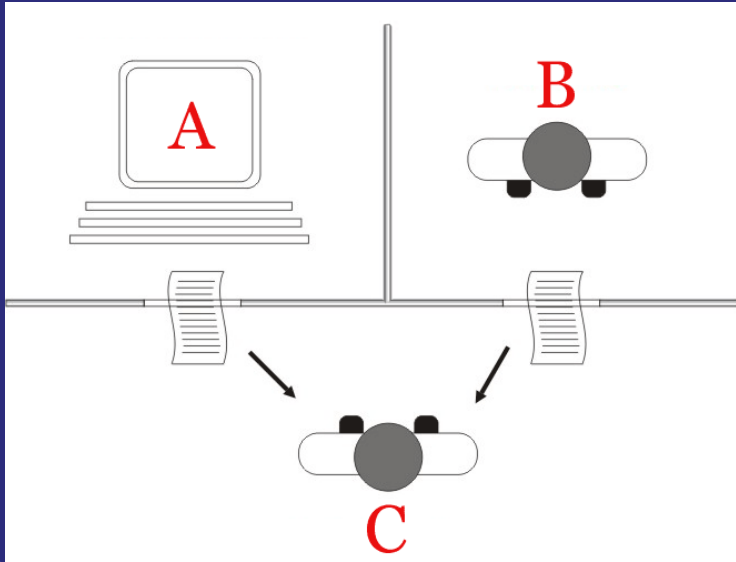


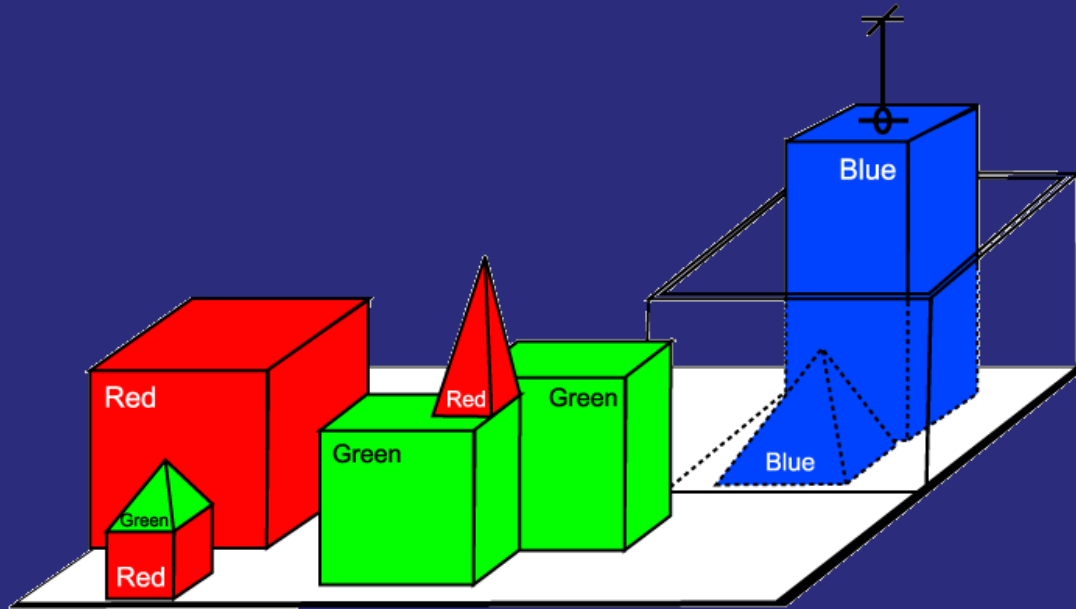
Illustration tirée de Wikipedia : « Test de Turing »

1950 - *Computing Machinery and Intelligence*

- Confrontation verbale à l'aveugle entre un examinateur humain (C) et une machine (A) ou un humain (B).
- Si l'examineur humain ne parviens pas à distinguer la machine d'un humain, la machine a passé le test.
- L'échange se fait à l'écrit.

Période symbolique

SHRDLU



Période symbolique

ELIZA

Psychothérapeute rogérien

Welcome to

```
EEEEEE LL      IIII ZZZZZZZ AAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LL      II      ZZZ  AAAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LLLLLL IIII ZZZZZZZ AA  AA
```

Eliza is a mock Rogerian psychotherapist.

The original program was described by Joseph Weizenbaum in 1966.

This implementation by Norbert Landsteiner 2005.

ELIZA: Please tell me what's been bothering you.

YOU: The weather is awful.

ELIZA: I'm not sure I understand you fully.

YOU: I hate the current weather.

ELIZA: You say you hate the current weather ?

YOU: Yes, that's what I said.

ELIZA: I understand.

YOU: Great.

ELIZA: Please go on.

YOU: I would say it's your turn now!

ELIZA: Why are you concerned over my turn now ?

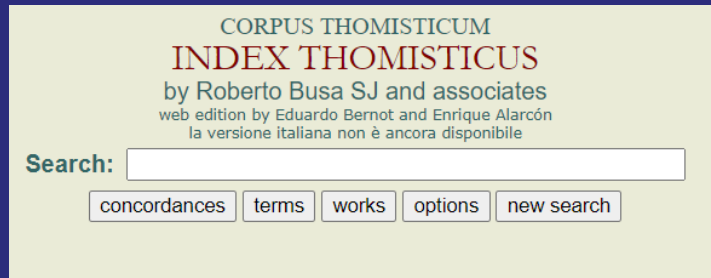
YOU: █

Période symbolique

Premières expériences en Linguistique Computationnelle

Padre Roberto Busa et l'*Index Thomisticus*

- Exploiter l'ordinateur comme moyen de archivage et analyse des données textuels
- Réaliser une vérification intégrale et ponctuelle du lexique de St Thomas d'Aquin, mais les analyses qu'il voulait faire n'étaient pas disponibles dans les concordanciers papier disponibles à l'époque
- Décide de réaliser sa propre base de données
- 194 - Demande à Thomas Watson (IBM) de lui donner des calculateurs
- En 1980, s'achève son travail, et publie les 70 000 pages de *l'Index Thomisticus* (11 millions de mots!)



<https://www.corpusthomisticum.org/it/index.age>

Index Thomisticus

Lemma 06079

anima: anima animae

Type of lemma: (A) common word

Type of meaning: (NV) invisible thing

cases	c. freq.	places	pl. freq.		form	type	infl.	num.	gen.	case	deg.	voice	tense	mood	pers.	comp.	notes
12983	1,22%	6780	7,78%	<input checked="" type="checkbox"/>	anima	n.	irr. decl.	sing.	f.	nom.	pos.						<i>b</i>
8511	0,80%	4921	5,65%	<input type="checkbox"/>	animae	n.	irr. decl.	sing.	f.	gen.	pos.						<i>g</i>
60	0,01%	26	0,03%	<input type="checkbox"/>	anime	n.	irr. decl.	sing.	f.	gen.	pos.						<i>b</i>
1	0,00%	1	0,00%	<input type="checkbox"/>	animaeque	n.	irr. decl.	sing.	f.	gen.	pos.					-que	
4325	0,41%	2694	3,09%	<input type="checkbox"/>	animam	n.	irr. decl.	sing.	f.	acc.	pos.						
1	0,00%	1	0,00%	<input type="checkbox"/>	animamque	n.	irr. decl.	sing.	f.	acc.	pos.					-que	
609	0,06%	419	0,48%	<input type="checkbox"/>	animarum	n.	irr. decl.	pl.	f.	gen.	pos.						
349	0,03%	278	0,32%	<input type="checkbox"/>	animabus	n.	irr. decl.	pl.	f.	dat.	pos.						
743	0,07%	547	0,63%	<input type="checkbox"/>	animas	n.	irr. decl.	pl.	f.	acc.	pos.						<i>b</i>

(*b*) Not divided: Homographs belonging to other lemma entries have been assigned to this form, awaiting the analysis of all of its occurrences.

(*g*) Base-form of subsequent graphic variants. (The choice of this word as a base-form does not conform to scientific, but to technical criteria.)

(*b*) Secondary graphic variant.

General notice: Homographs within the same lemma entry have not yet been divided.



ATALA

Association pour le Traitement Automatique des Langues

- Fondée en **1959** !
- Aide à l'organisation de la conférence TALN et de sa session RECITAL
- Une des premières sociétés savantes au monde à s'occuper de TAL (reconnu par ACL)



Association for
Computational Linguistics

- Fondée en **1962**
- Organise la conférence ACL annuelle
- Sponsor de la revue Computational Linguistics

HMM Hidden Markov Models

- Type d'algorithme utilisé pour étiqueter des séquences dès les années '80
- « An HMM is a probabilistic sequence model: given a sequence of units (words, letters, morphemes, sentences, whatever), it computes a probability distribution over possible sequences of labels and chooses the best label sequence » (Jurafsky & Martin, 2024, chap. 17, p. 8)

Markov Assumption: $P(q_i = a | q_1 \dots q_{i-1}) = P(q_i = a | q_{i-1})$

- *Markov assumption* : quand on prédit le futur, le passé n'est pas important, seulement le présent est pris en compte
- Une chaîne de Markov cachée permet de prendre en considération les événements observés (les mots d'un texte) ainsi que ceux cachés (les POS tags par exemple)

HMM (2)

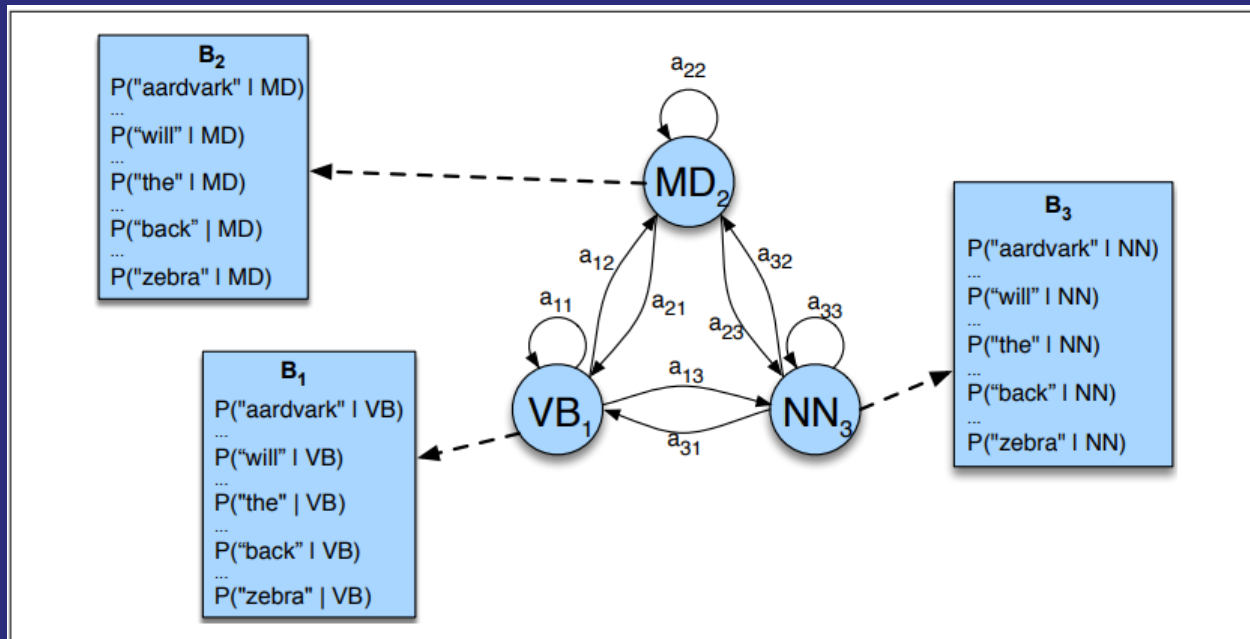


Figure 17.9 An illustration of the two parts of an HMM representation: the A transition probabilities used to compute the prior probability, and the B observation likelihoods that are associated with each state, one likelihood for each possible observation word.

Période statistique

1993-2012 ca

- Disponibilité d'une grande quantité de données langagières
- Augmentation de la puissance de calcul
- Développement des corpus multilingues et de l'annotation de corpus
- Introduction des algorithmes de Machine Learning

Applications :

- IBM Alignment Models (90s)

Période statistique

1993-2012 ca

2003 – le perceptron multi-niveau (Bengio et al., 2003)
obtient des meilleurs résultats du modèle word *n-gram*
2010 – RNN de Mikolov → Word2vec, word embeddings

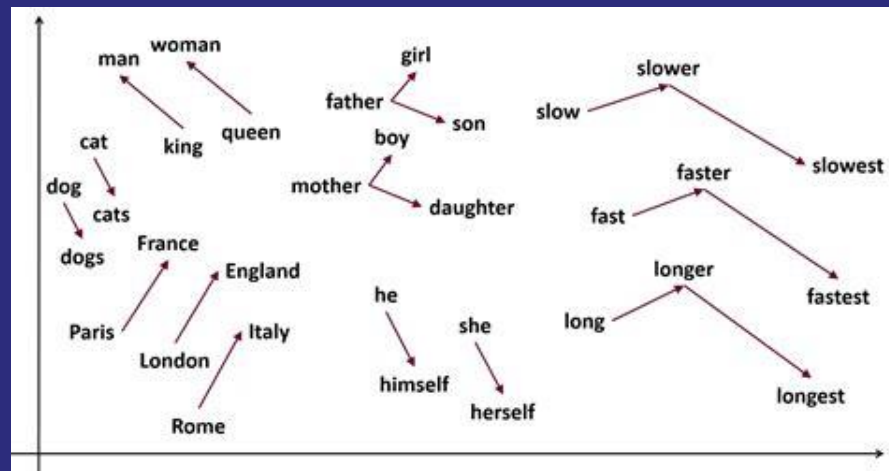
Mais aussi

• Treebanks et annotation de corpus

Word Embeddings

Intuition 1 - Chaque mot est associé à une composition de facteurs

Intuition 2 - Deux mots proches dans l'espace vectoriel partagent souvent des contextes similaires



Ex : le ... griffe ; ... est un félin

$occurrence(chat) \sim occurrence(tigre)$

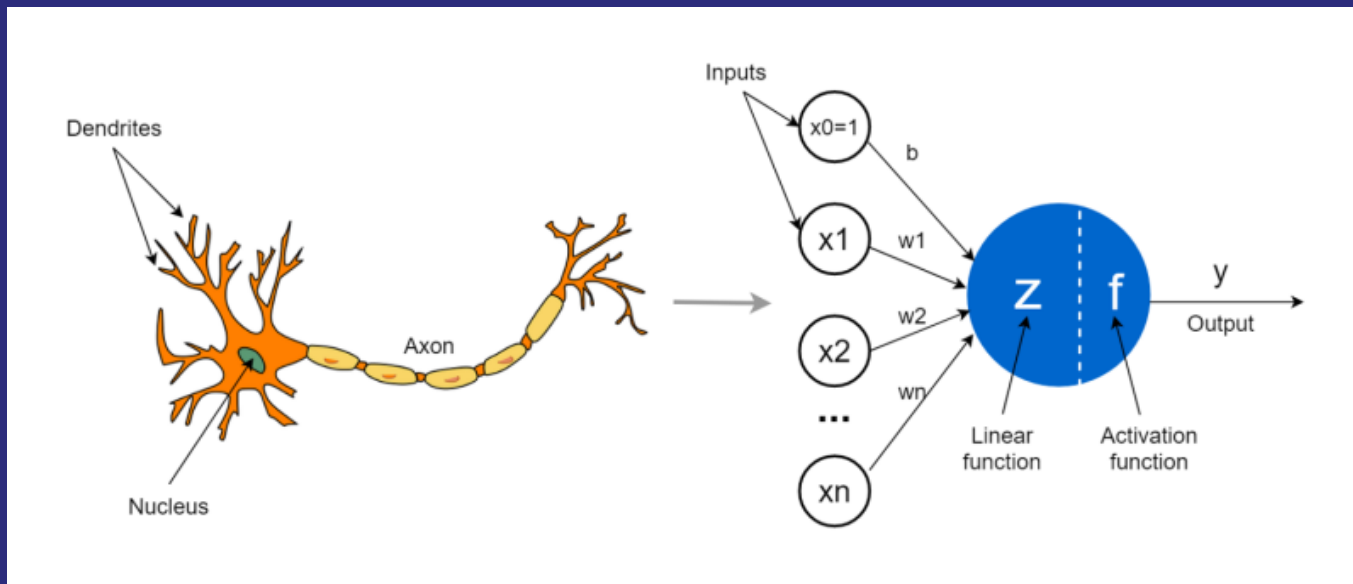
$w_{chat} \cdot w_{contexte} \sim w_{tigre} \cdot w_{contexte}$

$w_{chat} \sim w_{tigre}$

Période statistique

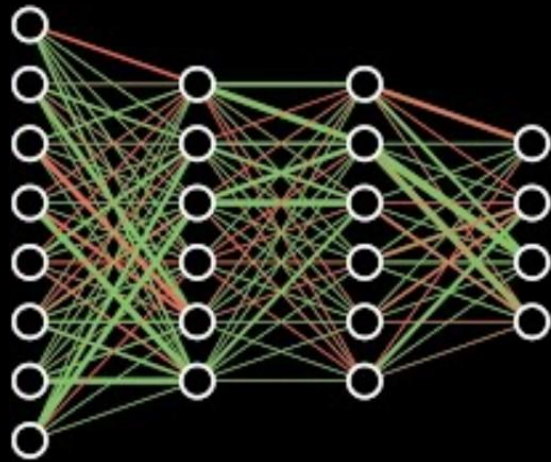
Le perceptron de Rosenblatt

- Classifieur linéaire



Bonus : What is a neural network?

Neural Networks



From the
ground up

Treebanks

Treebank – corpus parsé et annoté avec des étiquettes syntaxique et/ou en structures sémantiques (POS tag, analyse en dépendance)

Aujourd'hui...

Universal Dependencies

vs.

Surface Syntactic Universal Dependencies SUD

Universal Dependencies

- Un jeu d'étiquettes pour toutes les langues
- Le même format pour tous (CoNLL-U)

```
# sent_id = 1
# text = They buy and sell books.
1   They   they   PRON   PRP   Case=Nom|Number=Plur      2   nsubj   2:nsubj|4:nsubj   _
2   buy    buy    VERB   VBP   Number=Plur|Person=3|Tense=Pres 0   root     0:root           _
3   and    and     CCONJ  CC     _                        4   cc       4:cc             _
4   sell    sell    VERB   VBP   Number=Plur|Person=3|Tense=Pres 2   conj     0:root|2:conj    _
5   books   book    NOUN   NNS    Number=Plur              2   obj      2:obj|4:obj      SpaceAfter=No
6   .       .       PUNCT  .      _                        2   punct    2:punct          _

# sent_id = 2
# text = I have no clue.
1   I       I       PRON   PRP   Case=Nom|Number=Sing|Person=1    2   nsubj    _   _
2   have    have    VERB   VBP   Number=Sing|Person=1|Tense=Pres 0   root     _   _
3   no      no      DET    DT     PronType=Neg                  4   det      _   _
4   clue    clue    NOUN   NN     Number=Sing                   2   obj      _   SpaceAfter=No
5   .       .       PUNCT  .      _                        2   punct    _   _
```

Annotation de corpus

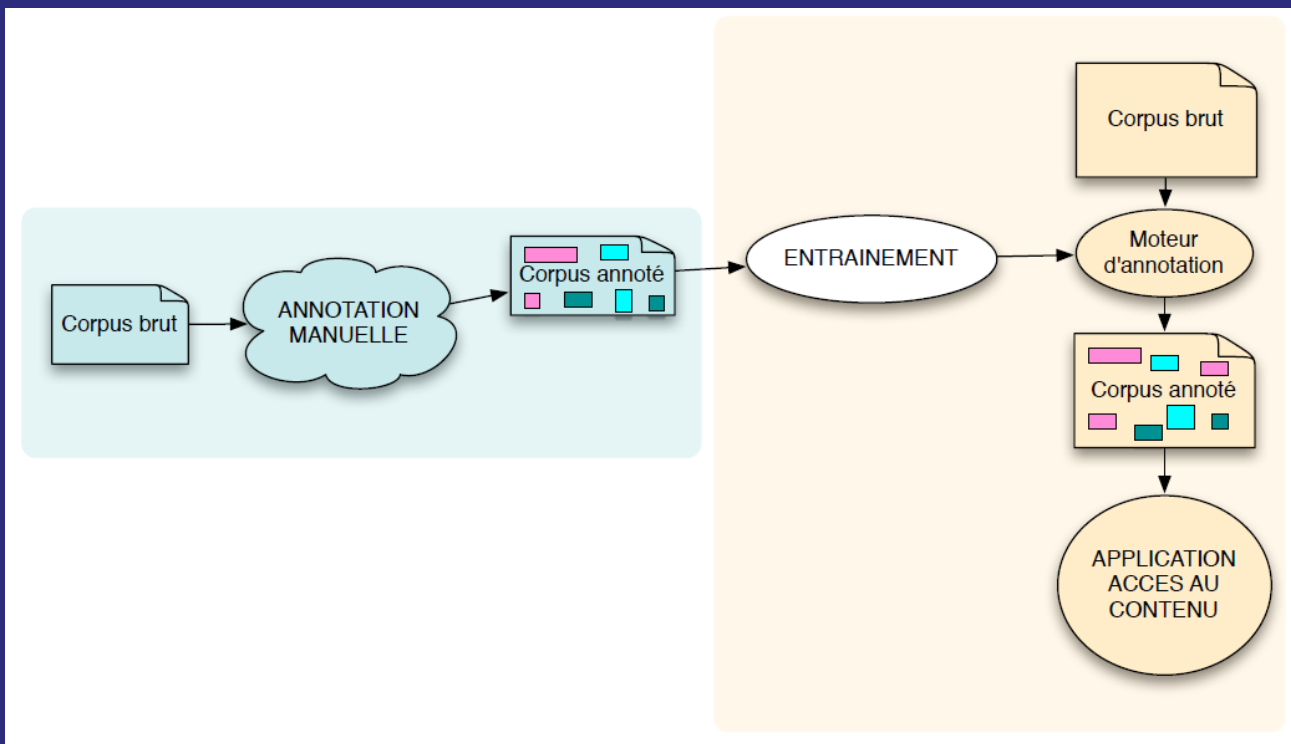


Figure tirée des diapos du cours de K. Fort, école d'été AnnoDemo 2024

Période neuronale

2013-aujourd'hui

- Mais déjà présentes en 2000 – Bengio et al., 2000
NLM basés sur LSTM ou GRU déjà utilisés en traduction automatique, génération et classification de textes

Période neuronale

2013-2017

- Différentes architectures se sont affirmées puis ont été surpassées par la suivante (CNN, LSTM, GRU, seq2seq...)
- Architecture Transformer

Vaswani et al., « Attention is all you need », 2017

2018-2023

- Approche Transformer dominant
 - Modèles entraînés sur une masse de données puis spécialisés pour une tâche spécifique

Période neuronale

2023-2024

L'explosion des LLM et « the rise of small Language Models »

- Modèles qui ont moins d'hyperparamètres
- Modèles accessibles (tournent sur Google Colab)

Famille des modèles Llama (Meta AI)

- Fournir des modèles de taille différentes pour obtenir la meilleure performance avec un budget computationnel donné

<https://huggingface.co/blog/2023-in-llms>

ChatGPT

Ou l'explosion des LLM

- Chatbot et assistant virtuel
- Appartient à la famille des modèles GPT – Generative Pre-trained Transformer
- Fine-tuned pour la tâche de conversation à travers apprentissage supervisé et reinforcement learning à travers feedback humain

- Des travailleurs payés 2\$ au Kenya pour épurer le côté « toxique » de ChatGPT

“Despite the foundational role played by these data enrichment professionals, a growing body of research reveals the precarious working conditions these workers face,” says the Partnership on AI, a coalition of AI organizations to which OpenAI belongs. “This may be the result of efforts to hide AI’s dependence on this large labor force when celebrating the efficiency gains of technology. Out of sight is also out of mind.”

(Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic, Time, 18 Janvier 2023)

ChatQPT

Ou l'explosion des LLM

“(...) that for all its glamor, AI often relies on hidden human labor in the Global South that can often be damaging and exploitative. These invisible workers remain on the margins even as their work contributes to billion-dollar industries.”

(Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic, Time, 18 Janvier 2023)

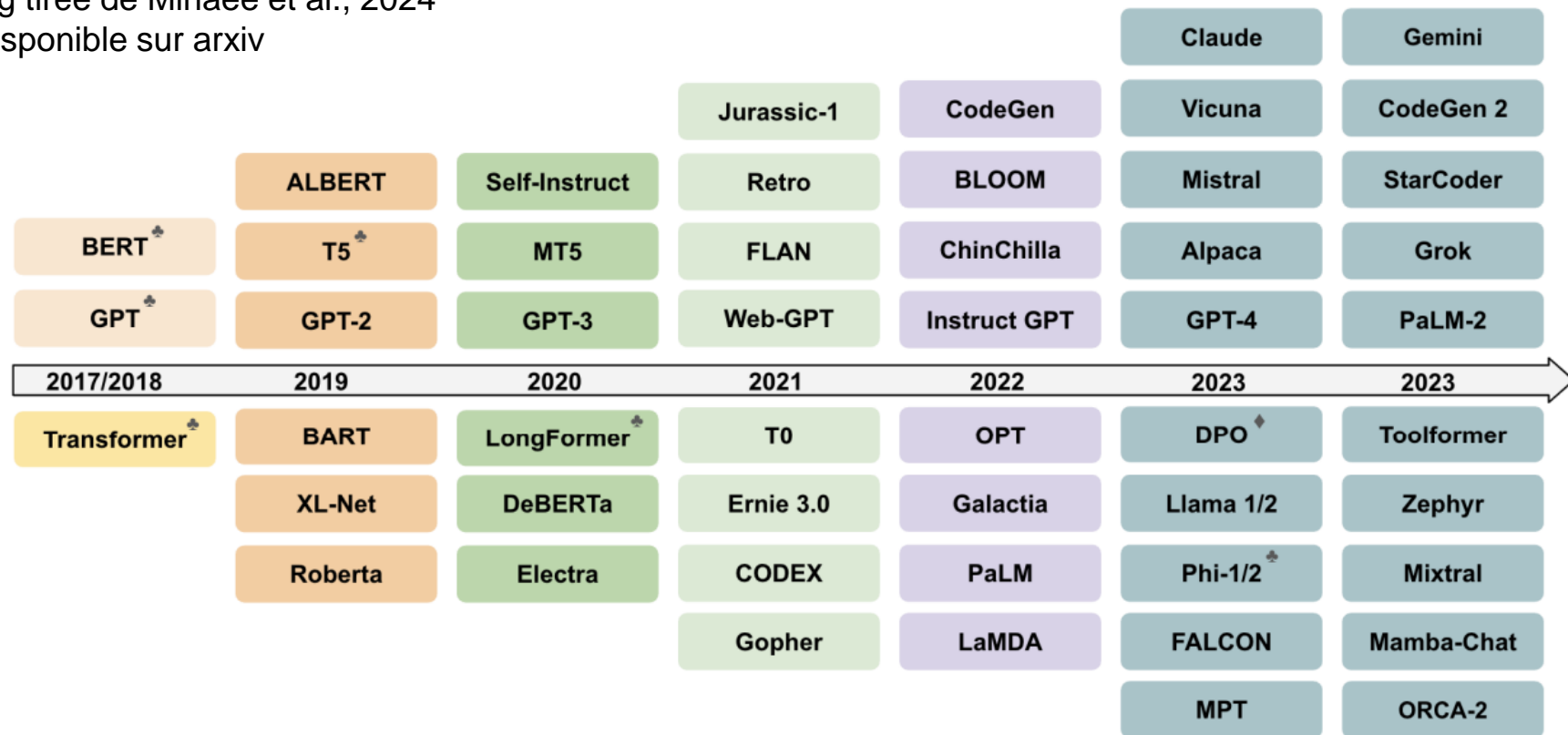


Fig. 24: Timeline of some of the most representative LLM frameworks (so far). In addition to large language models with our #parameters threshold, we included a few representative works, which pushed the limits of language models, and paved the way for their success (e.g. vanilla Transformer, BERT, GPT-1), as well as some small language models. ♠ shows entities that serve not only as models but also as approaches. ♦ shows only approaches.

Fig tirée de Minaee et al., 2024
Disponible sur arxiv

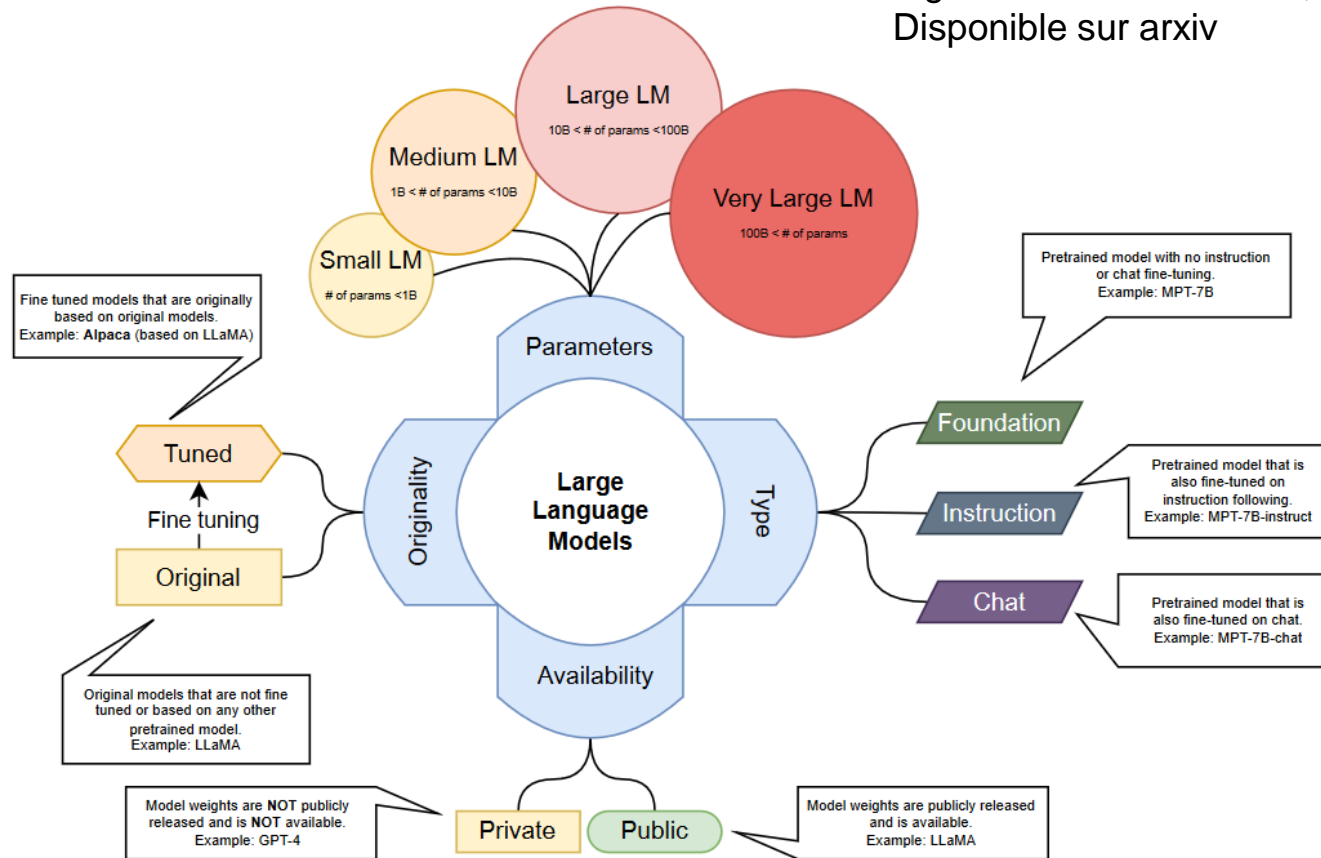


Fig. 43: LLM categorizations.

Fig tirée de Minaee et al., 2024
Disponible sur arxiv

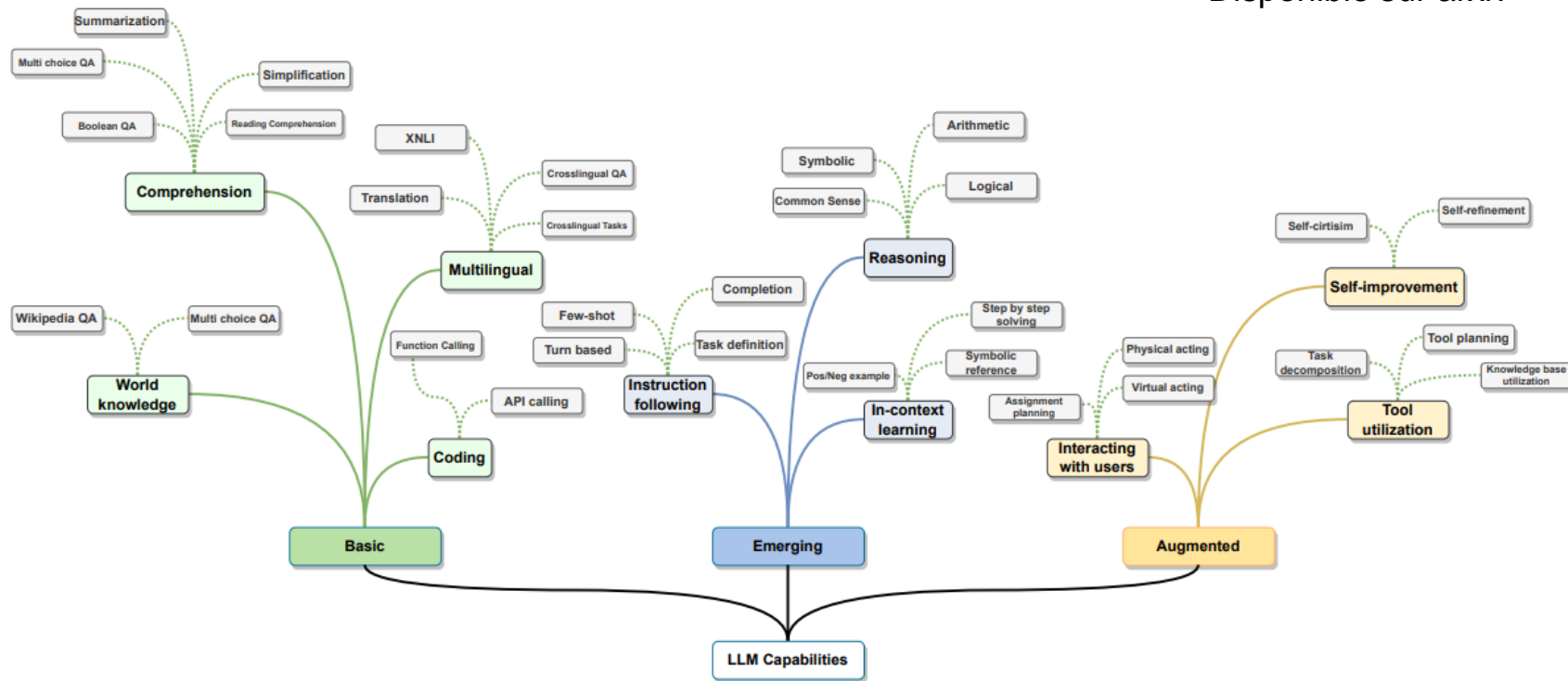
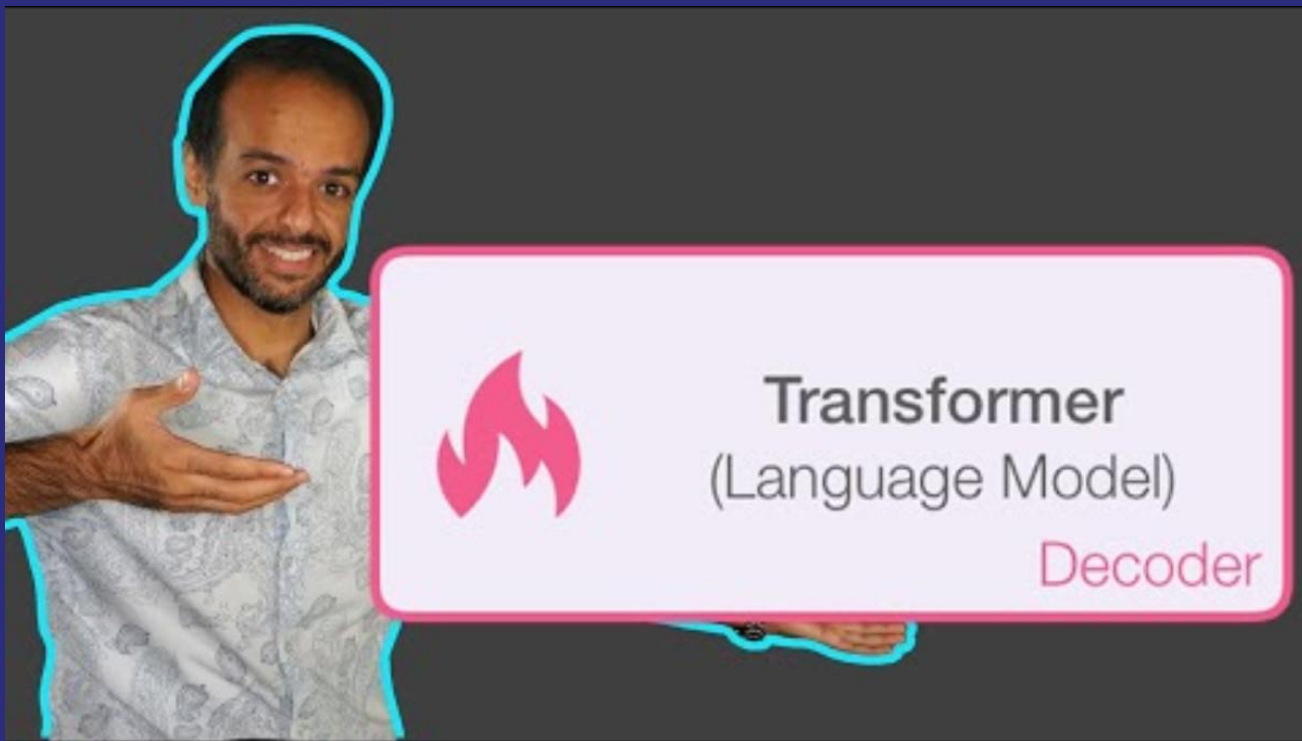


Fig. 1: LLM Capabilities.

Bonus : The Narrated Transformer



Bibliographie

- Alammar, J. (s. d.). *The Illustrated Transformer*. Consulté 3 septembre 2024, à l'adresse <https://jalammar.github.io/illustrated-transformer/>
- Chris Manning : *How computers are learning to understand language* | Stanford University School of Engineering. (2017, mai 22). <https://engineering.stanford.edu/magazine/article/chris-manning-how-computers-are-learning-understand-language>
- *Electronic brain translates from Russian to English*. (s. d.).
- Hutchins, J. (1995). « *The whisky was invisible* », or *Persistent myths of MT*. *MT News International*(11), 17-18.
- Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Vol. 2).
- Kerbrat-Orecchioni, C. (2005). *L'ambiguïté : Définition, typologie*. https://www.persee.fr/doc/mom_0184-1785_2005_act_33_1_2284
- Kraif, O. (s. d.). *Corpus parallèles, corpus comparables : Quels contrastes ?*
- Léon, J. (2015). La traduction automatique comme technologie de guerre. In *Histoire de l'automatisation des sciences du langage*. ENS Éditions. <https://doi.org/10.4000/books.enseditions.3737>
- Li, X. (2023). "There's No Data Like More Data" : Automatic Speech Recognition and the Making of Algorithmic Culture. *Osiris*, 38, 165-182. <https://doi.org/10.1086/725132>
- Manning, C. D. (2022). Human Language Understanding & Reasoning. *Daedalus*, 151(2), 127-138. https://doi.org/10.1162/daed_a_01905
- Morgan, N., & Boulard, H. (1990). Continuous speech recognition using multilayer perceptrons with hidden Markov models. *International Conference on Acoustics, Speech, and Signal Processing*, 413-416. <https://doi.org/10.1109/ICASSP.1990.115720>
- *Speech and Language Processing*. (s. d.). Consulté 23 août 2024, à l'adresse <https://web.stanford.edu/~jurafsky/slp3/>
- *Summary of the tokenizers*. (s. d.). Consulté 3 septembre 2024, à l'adresse https://huggingface.co/docs/transformers/tokenizer_summary
- *What is Perceptron? A Beginners Guide for 2023* | Simplilearn. (2021, mai 26). Simplilearn.Com. <https://www.simplilearn.com/tutorials/deep-learning-tutorial/perceptron>