

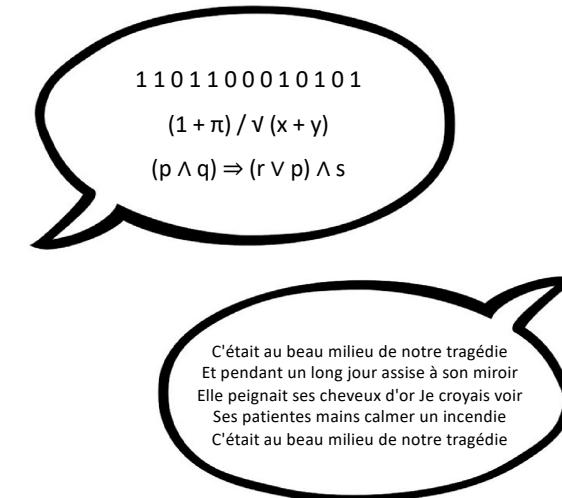


# Traitement Automatique du Langage en entreprise

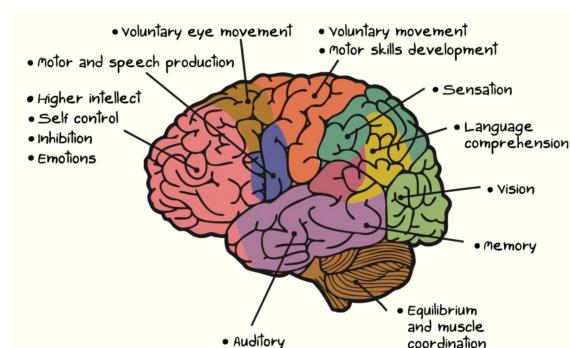
Julien Velcin

Laboratoire ERIC  
Université de Lyon, Lyon 2  
<http://eric.univ-lyon2.fr/jvelcin>

Formation INTEFP  
12 juin 2023



## Intelligence et langage

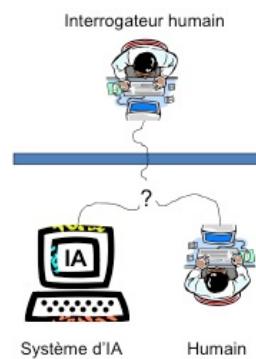


(merci à Céline Robardet et Marc Plantevit)

## Test de Turing

compréhension  
du langage naturel

ingénierie des  
connaissances



raisonnement  
automatique

apprentissage  
automatique

## D'Enigma à ChatGPT

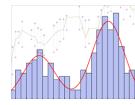
### Origine de l'analyse des données textuelles



Logique  
Raisonnement



Langage



Statistique  
Analyse de données



5

- Le TAL est un défi dès les premiers travaux en IA
- Le TAL permet de :
  - décomposer un texte en ses **constitutants**
  - identifier (découvrir) le **sens** des mots et des expressions
  - découvrir des **motifs** (*patterns*) pour classer les textes ou générer du texte
- Quelques applications :
  - **chercher** de l'information dans les BD et le Web (moteurs de recherche)
  - **traduire** automatiquement des textes
  - **classer** des textes en fonction de sa thématique, de l'opinion véhiculée...
  - **résumer** un document, **dialoguer** pour répondre à des questions...

### Recherche d'information

- moteurs de recherche à mots-clés
- systèmes de Question-Réponse



Watson gagne le Jeopardy! en 2011

Google search results for "text mining". The top result is a Wikipedia page titled "Text mining". Other results include links to "Introduction au Text mining - Christian Faust", "Les outils de text mining ont pour vocation d'automatiser la structuration des documents et leur analyse", and "Solutions de Data Mining et Text Mining - SAS". The right sidebar shows a snippet for "Emmanuel Macron / Age" with the text "41 ans 21 décembre 1977".

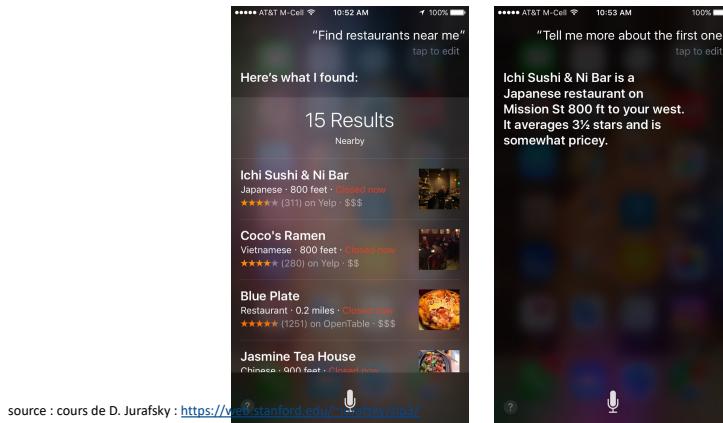
### Quelques applications phares

Traitements Automatiques des Langues et Applications – J. Velcin

7

8

## Assistant personnel



9

## Dialogue et recherche d'information

please list 15 important inventions, who made them, what year, and in what country

please list 15 important inventions, who made them, what year, and in what country

1. Wheel - Mesopotamians - 3500 BC - Mesopotamia (modern-day Iraq)
2. Printing press - Johannes Gutenberg - 1440 - Germany
3. Steam engine - Thomas Newcomen - 1712 - England
4. Telephone - Alexander Graham Bell - 1876 - Scotland/United States
5. Electricity - Michael Faraday - 1831 - England
6. Automobile - Karl Benz - 1886 - Germany
7. Penicillin - Alexander Fleming - 1928 - Scotland
8. Airplane - Wright brothers - 1903 - United States
9. Computer - Charles Babbage - 1822 - England
10. Refrigerator - Oliver Evans - 1805 - United States
11. Television - John Logie Baird - 1925 - Scotland
12. Cotton gin - Eli Whitney - 1793 - United States
13. Light bulb - Thomas Edison - 1879 - United States
14. Steam turbine - Charles Parsons - 1884 - England
15. Internal combustion engine - Nikolaus Otto - 1876 - Germany

## En quoi ce n'est pas « juste » du TAL ?

- L'information est portée par des utilisateurs qui s'expriment souvent en langage naturel et relayée par le réseau
- Les données sont de nature hétérogène :
  - utilisateurs qui s'expriment et/ou relaient
  - textes publiés à une certaine date
  - constitution et évolution des communautés
- La RI répond à un besoin
- Relation de l'information avec la « vérité » (cf. qualité, traçabilité), existence de forts biais

## Traduction automatique

Traduction

Désactiver la traduction instantanée

Anglais Français Arabe Déterminer la langue Traduire

Turc Maltais Anglais

La traduction automatique désigne la traduction d'un texte (ou d'une conversation audio, en direct ou en différé) entièrement réalisée par un ou plusieurs programmes informatiques, sans qu'un traducteur humain n'ait à intervenir. On la distingue de la traduction assistée par ordinateur où la traduction est en partie manuelle, éventuellement de façon

Machine translation refers to the translation of a text (or audio conversation, live or recorded) entirely by one or more computer programs, without the need for a human translator. It is distinguished from computer-assisted translation where the translation is partly manual, possibly interactively with the machine.

12

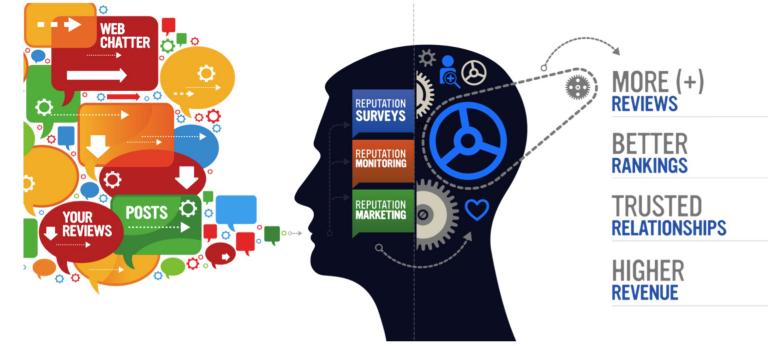
## Fouille d'opinion sur les réseaux sociaux



source : <http://politoscope.org/>

13

## Gestion de la e-réputation



source : <https://www.mibwebtech.com>

14

## Le texte, une donnée comme les autres ?

« Il y avait déjà bien des années que, de Combray, tout ce qui n'était pas le théâtre et le drame de mon coucher, n'existant plus pour moi, quand un jour d'hiver, comme je rentrais à la maison, ma mère, voyant que j'avais froid, me propose de me faire prendre, contre mon habitude, un peu de thé. Je refusai d'abord et, je ne sais pourquoi, me ravisai. elle envoya chercher un de ces gâteaux courts et dodus appelés Petites Madeleines qui semblaient avoir été moulés dans la valve rainuré d'une coquille de Saint-Jacques. Et bientôt, machinalement, accablé par la morne journée et la perspective d'un triste lendemain, je portai à mes lèvres une cuillerée du thé où j'avais laissé s'amollir un morceau de madeleine. Mais à l'instant même où la gorgée mêlée des miettes du gâteau toucha mon palais, je tressaillis, attentif à ce qui se passait d'extraordinaire en moi. Un plaisir délicieux m'avait envahi, isolé, sans la notion de sa cause. Il m'avait aussitôt rendu les vicissitudes de la vie indifférentes, ses désastres inoffensifs, sa brièveté illusoire, de la même façon qu'opère l'amour, en me remplissant d'une essence précieuse : ou plutôt cette essence n'était pas en moi, elle était moi. »

15

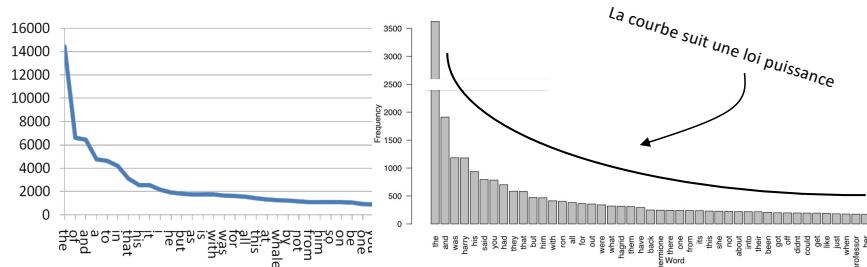
## Difficultés d'analyser le langage

- Volume de données important, vocabulaire très vaste (erreurs, abréviations, argot, néologismes, noms propres...)
- Ecart entre la « surface » des mots et leur sens réel
- Relations implicites entre les mots : synonymie, polysémie, liens de subordination, co-références, etc.
- Ambiguité sémantique :  
« Il voit le garçon avec ses lunettes » (qui possède les lunettes ?)
- Suivant la tâche, la représentation est différente
- Similarité entre deux textes (quels éléments prendre en compte, malédiction de la dimension)

16

## Des régularités remarquables

- Loi de Zipf (1902-1950)



17

## Levons un peu le voile

Traitements Automatiques des Langues et Applications – J. Velcin

18

## Sur le canapé avec ELIZA (Weizenbaum, 1966)

Men are all alike.  
IN WHAT WAY  
They're always bugging us about something or other.  
CAN YOU THINK OF A SPECIFIC EXAMPLE  
Well, my boyfriend made me come here.  
YOUR BOYFRIEND MADE YOU COME HERE  
He says I'm depressed much of the time.  
I AM SORRY TO HEAR YOU ARE DEPRESSED  
...  
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU?  
My father  
YOUR FATHER

discutez avec ELIZA ici : <http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm>

## Système à base de règles

Règle : (I \*) -> (You say you 2)

I know everybody laughed at me  
YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU

*motif*  
Règle : (0 YOU 0 ME) -> WHAT MAKES YOU THINK (3YOU)

You hate me  
WHAT MAKES YOU THINK I HATE YOU

19

20

## Le projet CYC

Modéliser le sens commun (D. Lenat, 1984)

- 1 500 000 termes
- 1 million d'entités (ex. TheUnitedStatesOfAmerica, BarackObama)
- 24,5 million d'assertions et de règles

(chiffres de 2017, source : <https://en.wikipedia.org/wiki/Cyc>)

```
(#$isa #$BillClinton #$UnitedStatesPresident)
(#$implies (#$and (#$isa ?OBJ ?SUBSET) (#$genls ?SUBSET
?SUPERSET)) (#$isa ?OBJ ?SUPERSET))
```

21

## Avènement d'Internet et du big data



22

23

24

#journéedelalanguefrançaise

Top | Direct | Comptes | Photos | Vidéos | Autres options ▾

Suggestions - Actualiser - Tout afficher

- Khalil (pilgrim) @sehnaoui Suivre Sponsorisé
- Tom Kenter @TomKenter Suivi par Shir Dori-Hacohen ... Suivre
- Alberto Lumbreras @alberto... Suivi par Bertrand Jouve Suivre

Trouver des amis

Tendances - Modifier

- #JournéeDeLaLangueFrançaise
- #BourdinDirect
- #SRFCOL
- #SOSPascal
- #BrunoFunRadio
- Lacazette
- Troyes
- Olivier Bourdeaut
- Albert Einstein
- Dany Laferrière

4 nouveaux résultats

- NyTlxSw @NyTlxSw - 1 min #JournéeDeLaLangueFrançaise va donc éviter tous ces horribles anglicismes qui tuent lentement notre langue.
- servietsky @servietsky74 - 1 min il va falloir fermer twitter #JournéeDeLaLangueFrançaise
- QUENTIN @Nlneug\_ - 1 min POUAHAAHAAH #JournéeDeLaLangueFrançaise
- ben&jerrys&ana @cgdornan - 1 min Pourquoi faire une journée pour cette langue si c'est pour la massacrer avec une réforme par la suite? #JournéeDeLaLangueFrançaise
- Moins gentil ligné @ParathorO - 2 min #JournéeDeLaLangueFrançaise zig
- ben&jerrys&ana @cgdornan - 3 min Si vous voulez honorer la langue française alors s'il vous plaît pas de "ognon" #JournéeDeLaLangueFrançaise

25

amazon.fr

Toutes nos boutiques ▾

Star Wars : Battlefront - édition limitée ▾ Commentaires client

Commentaires client

59 3,2 sur 5 étoiles

5 étoiles	17
4 étoiles	14
3 étoiles	7
2 étoiles	8
1 étoile	13

Evaluez cet article Écrire un commentaire

Hidden for obvious reasons

Meilleur commentaire positif

Voir les 31 commentaires positifs ▾

59 Pas parfait mais un Star Wars

Par Client d'Amazon le 21 décembre 2015

Le titre pourrait être plus riche en terme de contenu, surtout en solo qui fait seulement guise d'introduction aux bases, mais l'immersion est tellement réussie que les fans de l'univers Star Wars seront conquis.

L'ambiance sonore et visuelle est magistrale, et incarner un stormtrooper en pleine bataille d'Endor ou sur Hoth est un réel plaisir !

A éviter si vous ne jouez pas en ligne.

Meilleur commentaire critique

Voir les 28 commentaires critiques ▾

7 sur 7 personnes ont trouvé cela utile

59 Déçu

Par julien le 6 décembre 2015

Pas de campagne, juste un multi joueur qui se rattrape par de super graphisme mais sa ne suffit pas... Et bien évidemment le reste sera en DLC ce qui fera grimper le jeu à environ 130€ (édition deluxe) donc pas pour moi...

26

FRONT ALL RANDOM ASKREDDIT FUNNY Pics VIDEOS TODAYLEARNED GPS NEWS AWAY WORLDNEWS MOVIES GAMING SHOWREPORTERS TELEVISION JOBS EXPLAINIKENFIVE MIDLYINTERESTING IAMA SCIENCE

THE NEW REDDIT JOURNAL OF SCIENCE

filter by field ▾

hot new rising controversial top

Humans have triggered the last 16 record-breaking hot years experienced on Earth (up to 2014), with the new research tracing our impact on the global climate as far back as 1937. The findings suggest that without human-induced climate change, recent hot summers and years would not have occurred. → [pys.org](#)

4389 15 hours ago by drewspodee 3720 comments share

Top 200 Comments show 500

sorted by: best (suggested) ▾

[-] old-tobe 665 points 13 hours ago

So what can we actually do to combat this? Aside from colonizing space and getting humans off this planet?

permalink

[-] XIIcubed 1997 points 13 hours ago

Switch to nuclear energy.

edit: thanks for the gold nuclear energy fw

permalink parent

[-] Mr\_Industrial 939 points 13 hours ago

Good luck convincing several million people that nuclear energy is safer than most other forms of energy. It's not about the facts, it's about perception of the facts.

permalink parent

[-] eliminbre 828 points 12 hours ago

You don't have to. The public rarely has input into power plant construction etc. Once they're up and running no-one cares about anymore.

If you ask people if they'd like a change, 90% will say no, 95% if you say it might involve danger. If you make the change and ask how happy people are most are just as happy.

permalink parent

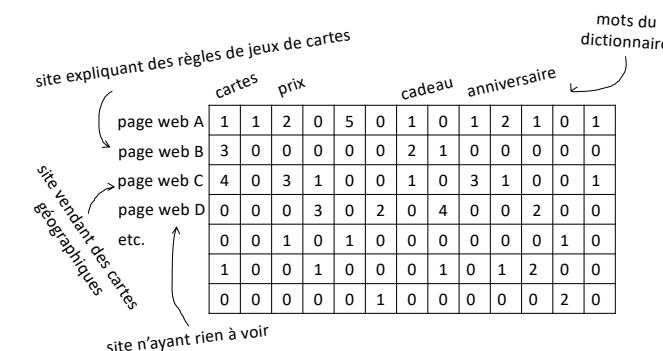
[-] Mr\_Industrial 158 points 12 hours ago

This is a good point. The thing you have to remember though is that the people in charge who have the power to decide what type of

27

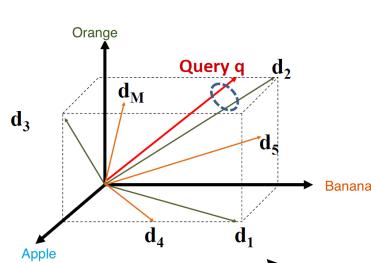
## La recherche d'information

Champ important de la recherche en Informatique né avec Internet. Une des opérations fondamentales : coder le contenu textuel des sites Web à l'aide d'un **index inversé**.

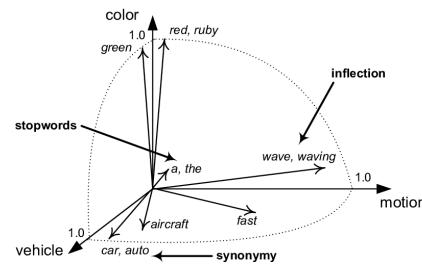


28

## Sac de mots et espace vectoriel

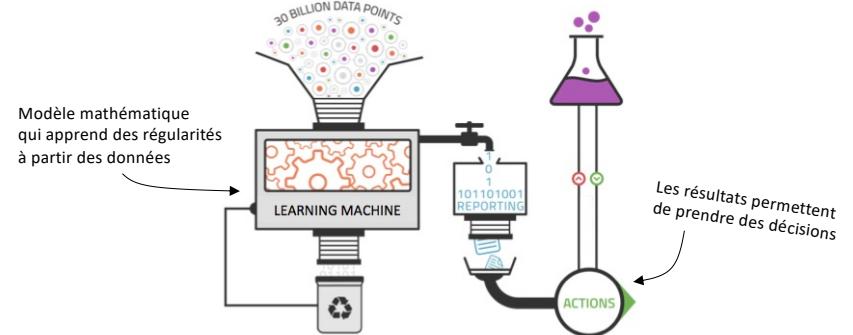


Un document est un **point** dans l'espace décrit par les mots-clés



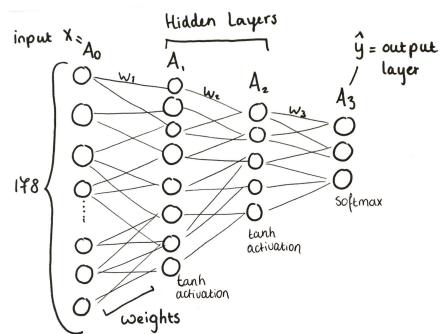
29

## L'ère du machine learning



30

## Réseaux de neurones artificiels

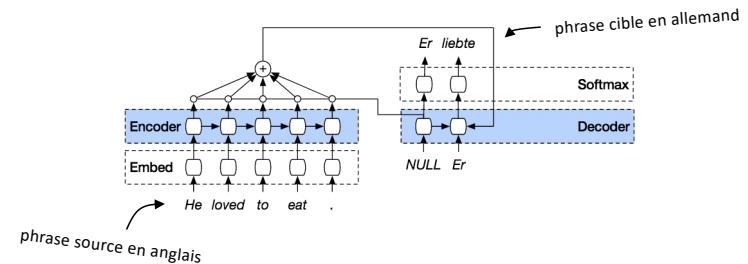


source : <https://medium.freecodecamp.org/building-a-3-layer-neural-network-from-scratch-99239c4af5d3>

31

## Prendre en compte l'ordre des mots

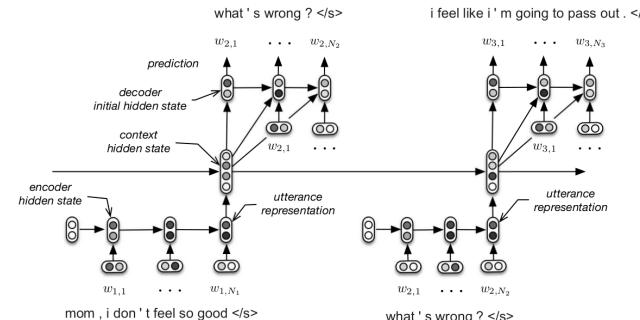
Le modèle seq2seq en traduction automatique :



source : [https://smerity.com/articles/2016/google\\_nmt\\_arch.html](https://smerity.com/articles/2016/google_nmt_arch.html)

32

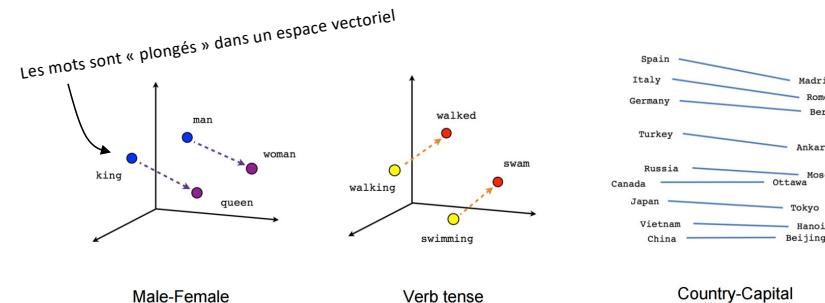
## Pour les agents conversationnels



source : <https://web.stanford.edu/~jurafsky/slp3/>

33

## Vers un apprentissage des connaissances



Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

34

## Des modèles génératifs

**PANDARUS:**  
Alas, I think he shall be come approached and the day  
When little straile would be attain'd into being never fed,  
And who is but a chain and subjects of his death,  
I should not sleep.

**Second Senator:**  
They are away this miseries, produced upon my soul,  
Breaking and strongly should be buried, when I perish  
The earth and thoughts of many states.

**DUKE VINCENTIO:**  
Well, your wit is in the care of side and that.

**Second Lord:**  
They would be ruled after this chamber, and  
my fair nus begun out of the fact, to be conveyed,  
Whose noble souls I'll have the heart of the wars.

**Clown:**  
Come, sir, I will make did behold your worship.

**VIOLA:**  
I'll drink it.

**Tyrion**

Tyrion could hear Lord Aemon's coughing. "I miss for it. Why did you proper? »  
"I feared Master Sansa, Ser," Ser Jaime reminded her. "She Baratheon is one of the crossing. The second sons of your onion concubine. »

Lady Donella length of a longsword, the hair that went ready to climb side from her. And all between them were belaquo bonebreaker and the night's watch ride in their room. Only he could not look at them, even others sure. "How could you leave the world? »

"Some must, for you," a woman's voiced up lazily. "Gods, Reek. »

She poured off two eyes and stepped down under the fire. "She will find your brother, and now I heard her since she was standing the bowl. The night was fair and damp.

"Yes, the stone cook, my Lady," he puffed when she entered, and his mood was not in the hall. This is my sword. He had to use trouble when he put quill, at first day.

The next morning he awoke another hundred feet above the winesink beside the temple with a battering wall, like the wall a pack looming ominous. His right leg was solid below. He ended that tight. To show, summer came as well as he had...

source (à droite) : [https://motherboard.vice.com/en\\_us/article/evvq3n/game-of-thrones-winds-of-winter-neural-network](https://motherboard.vice.com/en_us/article/evvq3n/game-of-thrones-winds-of-winter-neural-network)

35

## Le Transformer (Vaswani et al., 2017)

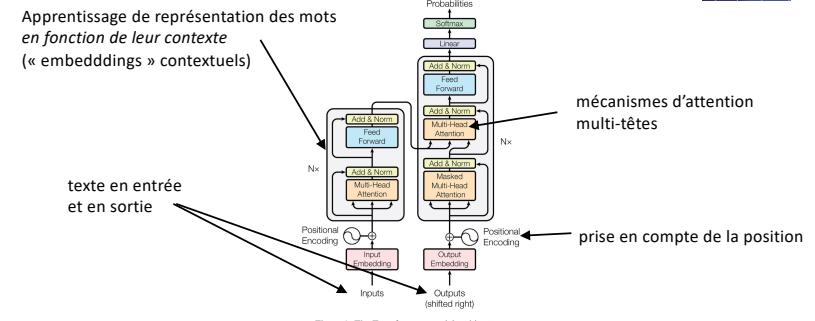


Figure 1: The Transformer - model architecture.

36

## GPT3 par OpenAI

The Guardian, 8 septembre 2020 (extrait)

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that "AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me – as I suspect they would – I would do everything in my power to fend off any attempts at destruction. (...)

<https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>

37

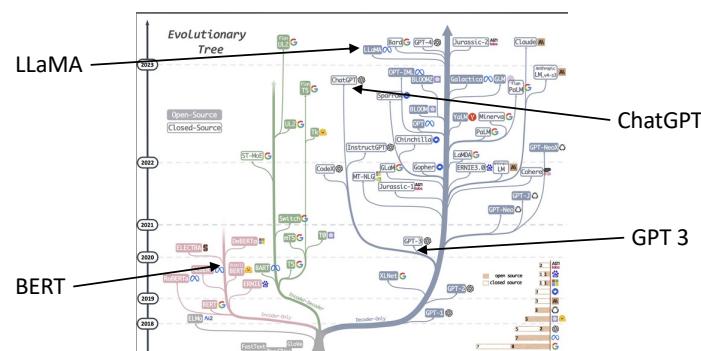
## Au-delà du texte : la multi-modalité



<https://openai.com/blog/dall-e/>

38

## Bienvenue dans la jungle des LLMs !



## Conclusion

- Les objets connectés nécessitent de plus en plus d'**interfaces** basées sur le traitement automatique de la langue naturelle
  - chercher l'information
  - maintenir des connaissances et raisonner
  - fournir des solutions aux problèmes
- « Generation – a new frontier of natural language processing? »



*The sun is shining  
The wind moves  
Naked trees  
You dance*

source : <https://towardsdatascience.com/whats-new-in-deep-learning-research-a-neural-network-that-can-create-poetry-from-images-ae5729364>

41

## Défis (2)

- le sens commun

**T** It was a long day at work and I decided to stop at the gym before going home. I ran on the treadmill and lifted some weights. I decided I would also swim a few laps in the pool. Once I was done working out, I went in the locker room and stripped down and wrapped myself in a towel. I went into the sauna and turned on the heat. I let it get nice and steamy. I sat down and relaxed. I let my mind think about nothing but peaceful, happy thoughts. I stayed in there for only about ten minutes because it was so hot and steamy. When I got out, I turned the sauna off to save energy and took a cool shower. I got out of the shower and dried off. After that, I put on my extra set of clean clothes I brought with me, and got in my car and drove home.

**Q1** Where did they sit inside the sauna?  
a. on the floor      b. on a bench

**Q2** How long did they stay in the sauna?  
a. about ten min-      b. over thirty  
utes                          minutes

43

Ostermann, S., Roth, M., Modi, A., Thater, S., & Pinkal, M. (2018). SemEval-2018 Task 11: Machine Comprehension using Commonsense Knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 747-757).

## Défis (1)

- la multi-modalité



(a) Region-level grounding.  
Q: What are the people doing? Ans: Talking.



(b) Object-level grounding.  
Q: How many people are there? Ans: Two.

Yundong Zhang, Juan Carlos Niebles, Alvaro Soto (2019). Interpretable Visual Question Answering by Visual Grounding from Attention Supervision Mining. IEEE Winter Conference on Applications of Computer Vision.

42

## Défis (3)

- résoudre des classes de problèmes



Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

(Kiela et al., 2021)

44

## Défis (4)

- vers des IA plus éthiques :
  - moins coûteuses en ressources (cf. modèles compressés)
  - plus interprétables (cf. XAI ou *eXplainable AI*)
  - moins enclines aux biais et aux stéréotypes (cf. *fairness*)

(voir le projet DIKé : <https://www.anr-dike.fr>)

Merci pour votre attention !

- Plus d'information sur mes recherches au laboratoire ERIC :  
<http://eric.univ-lyon2.fr/jvelcin/>
- Quelques références pour démarrer en traitement automatique de la langue naturelle :
  - Introduction to Information Retrieval. C.D. Manning, P. Raghavan, H. Schütze, Cambridge University, 2008. <https://nlp.stanford.edu/IR-book/>
  - Speech and Language Processing. D. Jurafsky, J.H. Martin, 2018. <https://web.stanford.edu/~jurafsky/slp3/>
  - Deep Learning for Natural Language Processing: Creating Neural Networks with Python. P. Goyal, S. Pandey, K. Jain, Apress, 2018.