université PARIS-SACLAY

ANALYSE DES DONNÉES TEXTUELLES & ÉCHANGES HOMME-MACHINE

29 mai 2024
GdR MaDICS, Blois

Vincent Guigue & Laure Soulier
vincent.guigue@agroparistech.fr, laure.soulier@isir.upmc.fr

MIA PARIS-SACLAY
EKINOCS

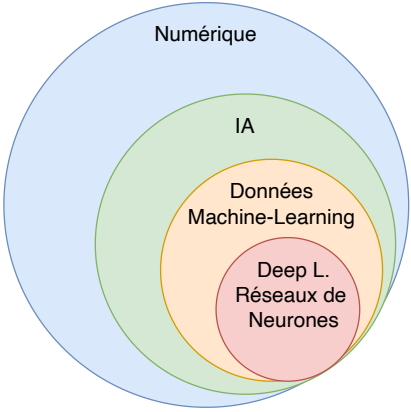
AGROPARISTECH
Institut des Sciences et Industries du Vivant et de l'Environnement

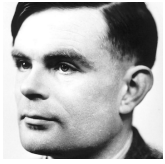
Introduction

Historique générale de l'Intelligence Artificielle

- Deux concepts distincts malgré les liens
- IA: Différentes Définitions

1956 N'importe quel algorithme / programme
1960-2012 Systèmes experts et raisonnement logique
2012- Données & réseaux de neurones





A. Turing

Marvin Minsky

G. Hinton

Y. Lecun

Ordinateur — Réseaux de neurones — Deep-learning

1941 — 1956 — 1986 — 2012

Informatique

IA: grande variété d'algorithme
Principalement : Système expert / raisonnement logique

IA = réseaux de neurones

Intelligence Artificielle & Machine Learning

Input (**X**)	Output (**Y**)	Application
email	spam? (0/1)	spam filtering
audio	text transcript	speech recognition
English	Chinese	machine translation
ad, user info	click? (0/1)	online advertising
image, radar info	position of other cars	self-driving car
image of phone	defect? (0/1)	visual inspection

IA : programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau.

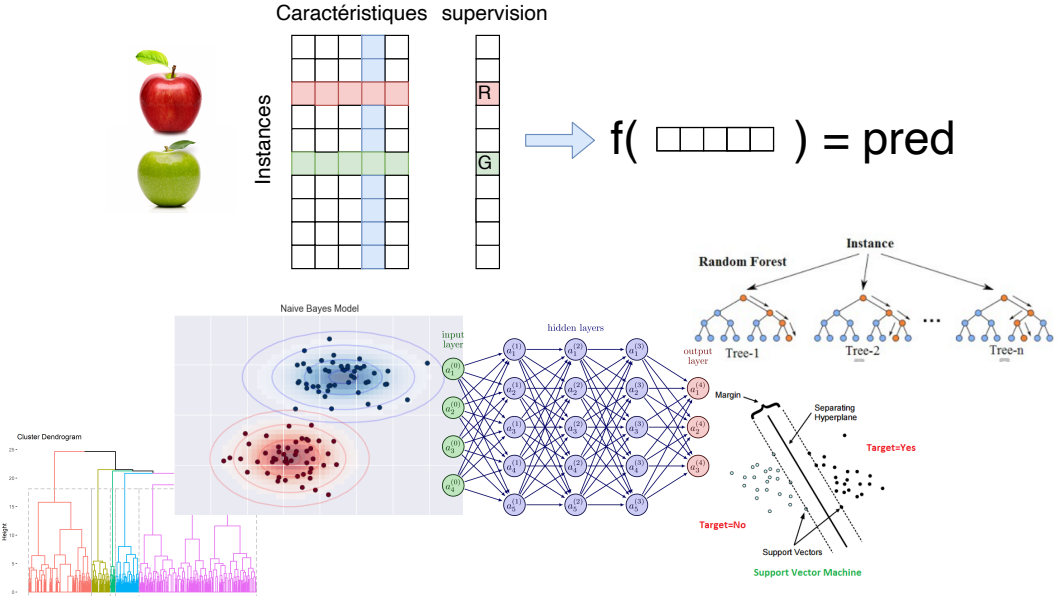
Marvin Lee Minsky, 1956

N-AI (Narrow Artificial Intelligence), dédiée à une tâche

$\neq$ **G-AI (General AI)** qui remplace l'humain dans des systèmes complexes.

Andrew Ng, 2015

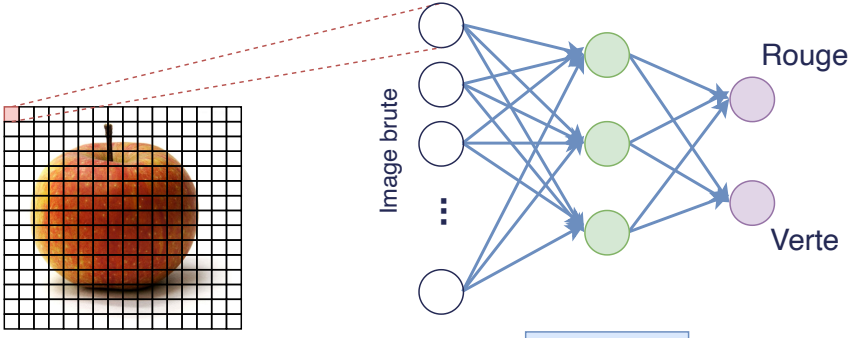
Big Data (2001): Le défi des données



Réseaux de neurones

Une fonction complexe & protéiforme ⇒ Adaptable à beaucoup de problèmes

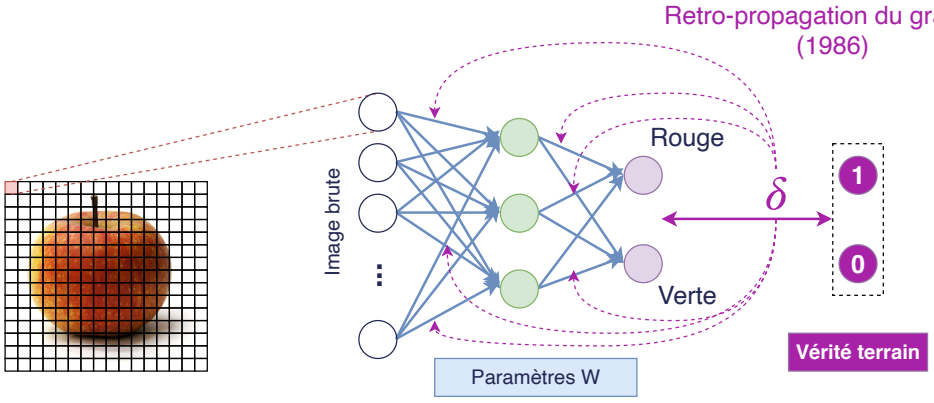
(1) Initialisation aléatoire (& comportement aléatoire)



Réseaux de neurones

Une fonction complexe & protéiforme ⇒ Adaptable à beaucoup de problèmes

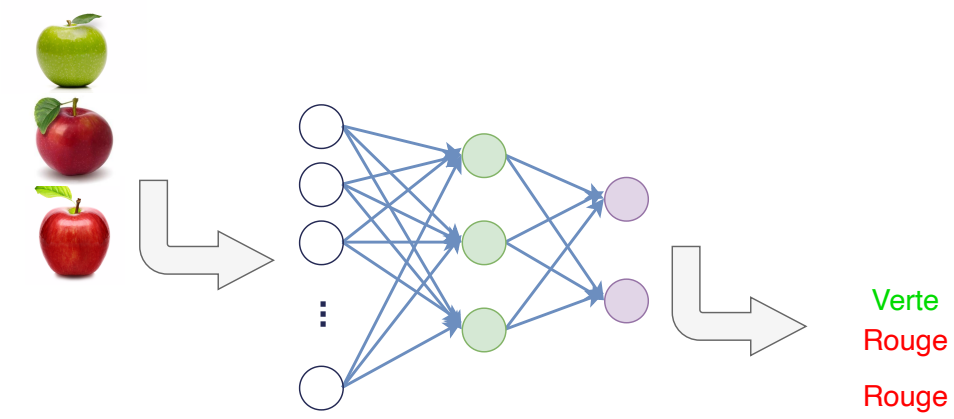
(2) Entrainement lent, long & stochastique



Réseaux de neurones

Une fonction complexe & protéiforme ⇒ Adaptable à beaucoup de problèmes

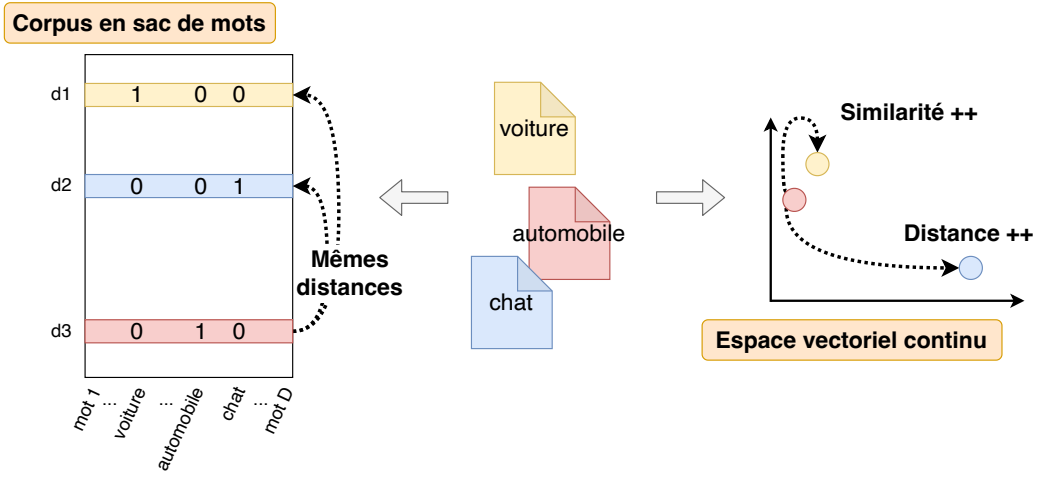
(3) Inférence rapide



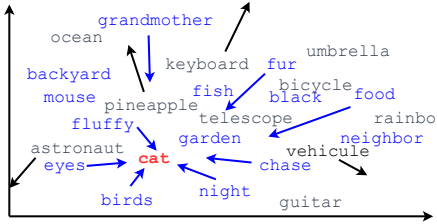
Deep-Learning ⇒ Representation Learning

Enjeu: l'apprentissage de représentation

Comprendre comment des concepts complexes (mot/objet/image) se positionnent les uns par rapport aux autres

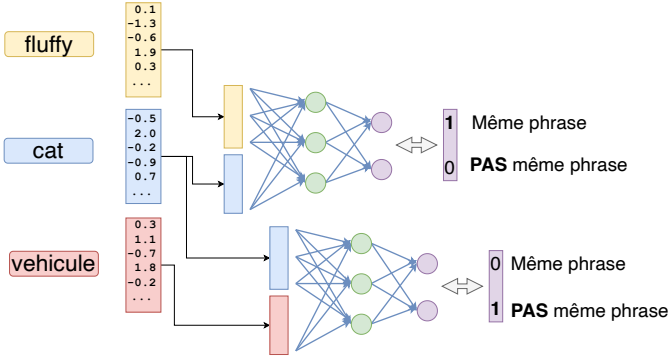


Fonctionnement type Word2Vec



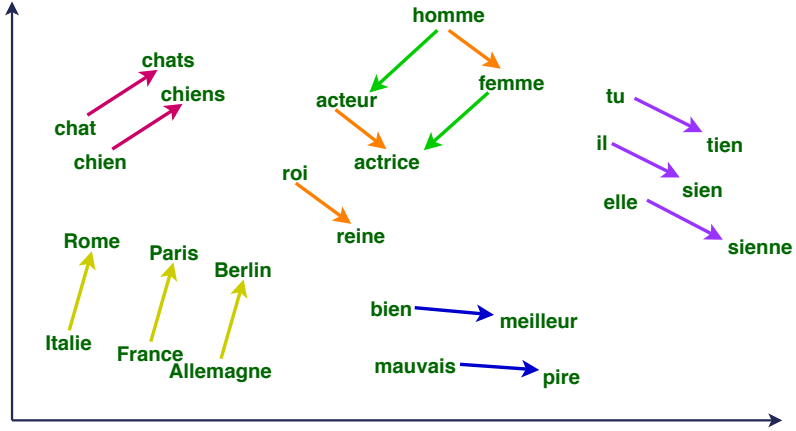

```
The fluffy cat napped lazily in the sunbeam.
I adopted a stray cat from the shelter last week.
My cat loves to chase after toy mice.
The black cat stealthily crept through the dark alley.
I often find my cat perched on the windowsill, watching birds.
She gently stroked her cat's fur as it purred contentedly.
Our neighbor's cat frequently visits our backyard.
The playful cat swatted at the dangling string with its paw.
My cat has a preference for fish flavored cat food.
The cat stealthily stalked a mouse in the garden.
My grandmother has a collection of porcelain cat figurines.
```

Corpus



Fonctionnement type Word2Vec

- Espace sémantique : signification proche ⇔ position proche
- Espace structuré : régularités grammaticales, logiques, ...

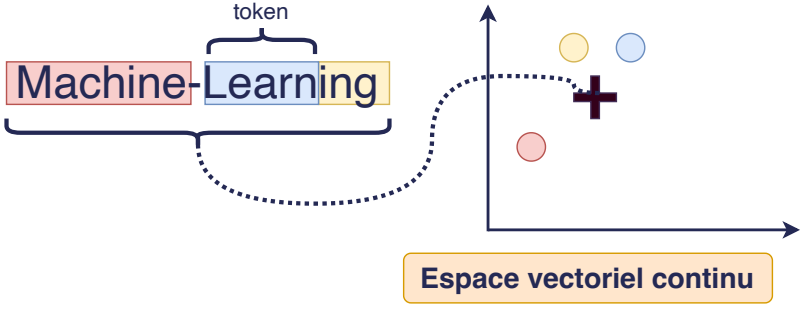


Distributed representations of words and phrases and their compositionality, Mikolov et al. NeurIPS 2013

Des mots aux tokens

- Problème de taille du dictionnaire / mots inconnus
- Résistance aux fautes d'orthographe

Décomposition en groupes de lettres fréquents

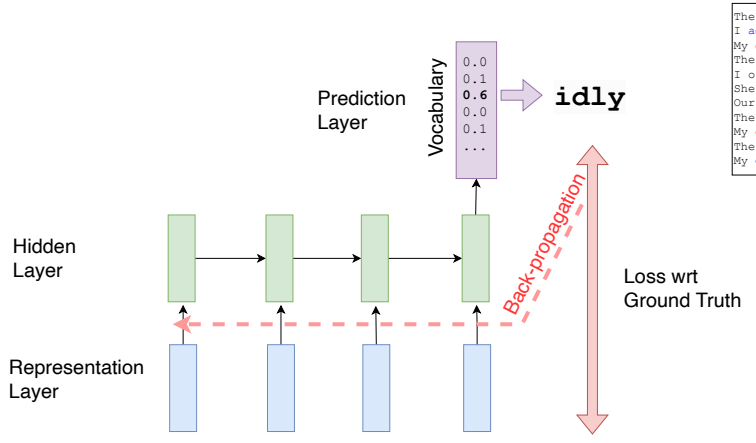


Enriching Word Vectors with Subword Information, Bojanowski et al. TACL 2017

Des modèles de langue à chatGPT

Vers les modèles de langue : Agrégation & Prédiction

- Nouvelle manière d'apprendre les positions des mots
- IA générative : traduction, résumé automatique



The fluffy cat napped lazily in the sunbeam.
I adopted a stray cat from the shelter last week.
My cat loves to chase after toy mice.
The black cat stealthily crept through the dark alley.
I often find my cat perched on the windowsill, watching birds.
She gently stroked her cat's fur as it purred contentedly.
Our neighbor's cat frequently visits our backyard.
The playful cat swatted at the dangling string with its paw.
My cat has a preference for fish flavored cat food.
The cat stealthily stalked a mouse in the garden.
My grandmother has a collection of porcelain cat figurines.

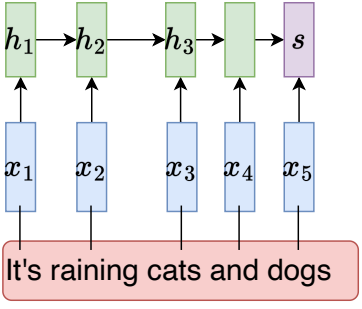
Corpus

Sequence to sequence learning with neural networksy, Sutskever et al. NeurIPS 2014

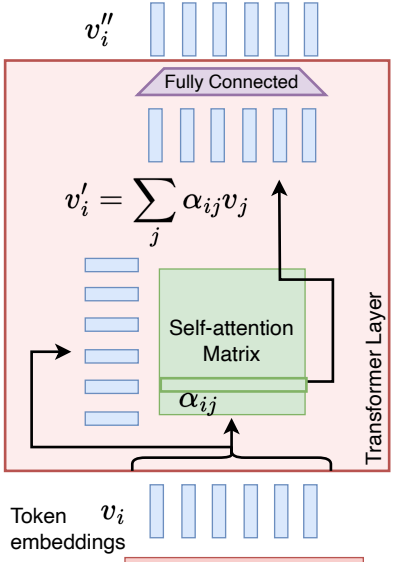
Passage aux Transformers

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:

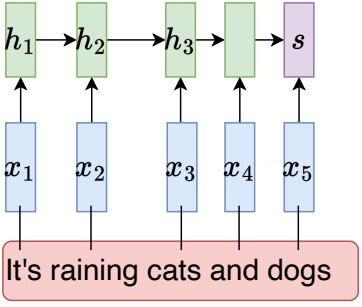


Attention is all you need, Vaswani et al. NeurIPS 2017

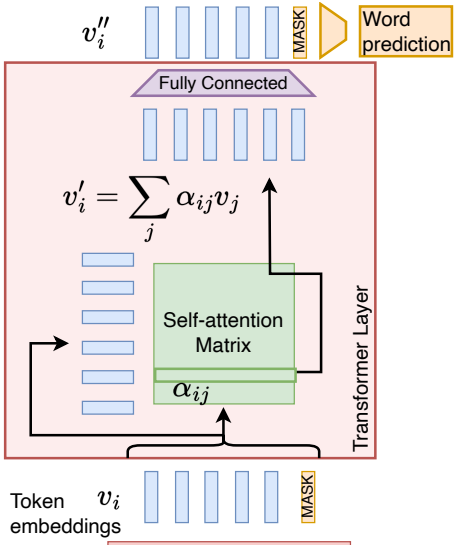
Passage aux Transformers

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:

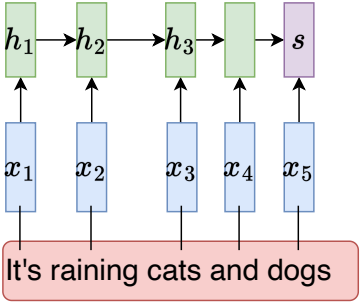


Attention is all you need, Vaswani et al. NeurIPS 2017

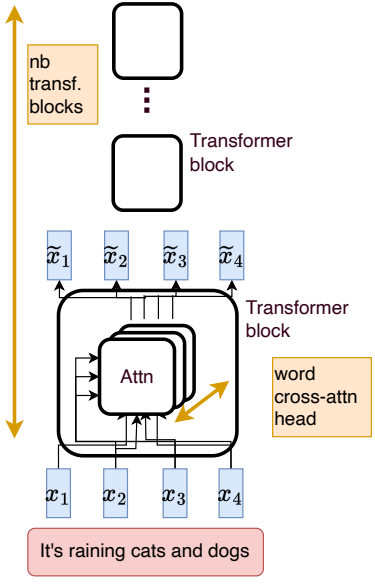
Passage aux Transformers

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:

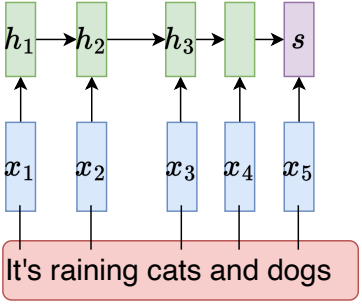


Attention is all you need, Vaswani et al. NeurIPS 2017

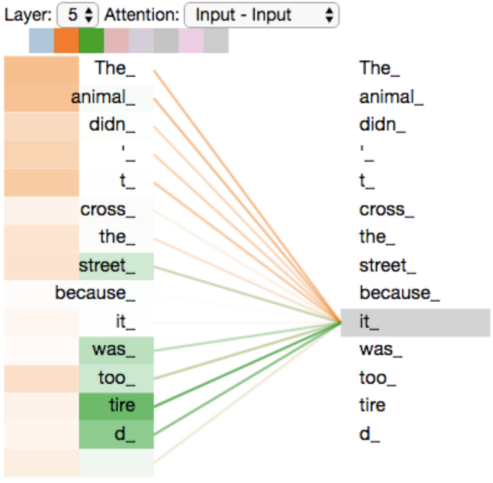
Passage aux Transformers

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$

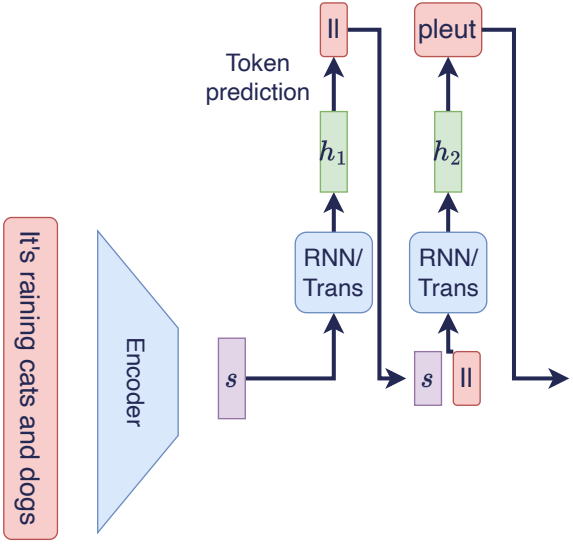


Transformer:



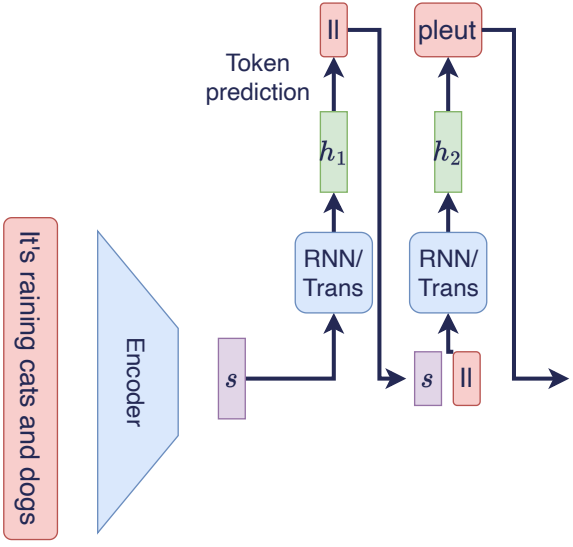
Attention is all you need, Vaswani et al. NeurIPS 2017

Architectures génératives / encodeur-décodeur

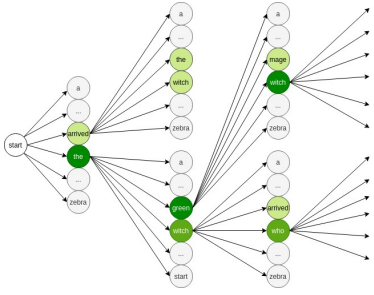


- Cout élevé (+beam search)
  - 1 appel / token
- Génération au sens du maximum de vraisemblance
- Principales tâches de NLP $\Leftrightarrow$ reformulation en mode génératif

Architectures génératives / encodeur-décodeur

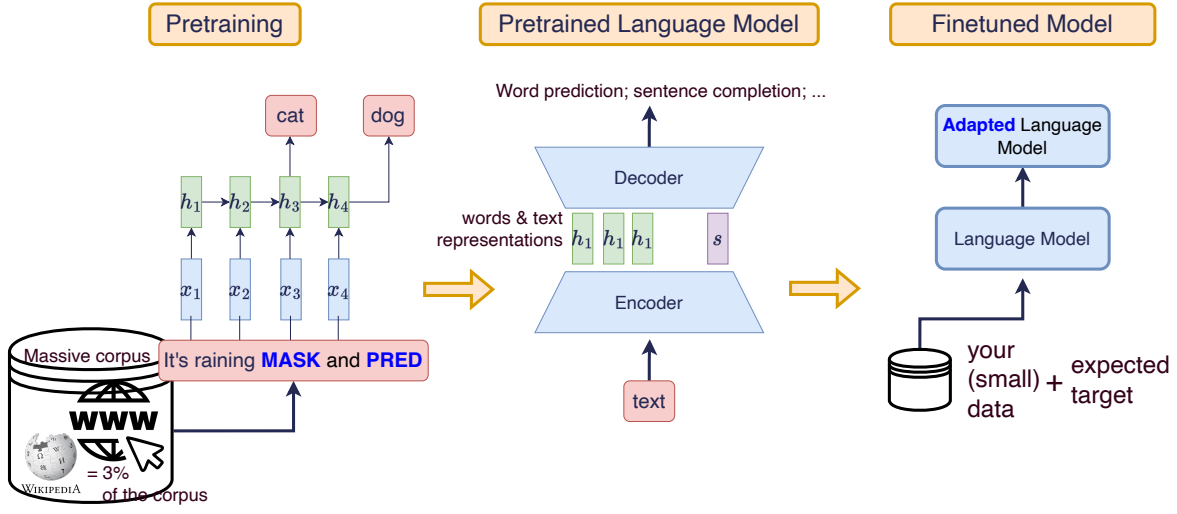


- Cout élevé (+beam search)
  - 1 appel / token
- Génération au sens du maximum de vraisemblance
- Principales tâches de NLP $\Leftrightarrow$ reformulation en mode génératif



Changement de paradigme: modèles pré-entrainés

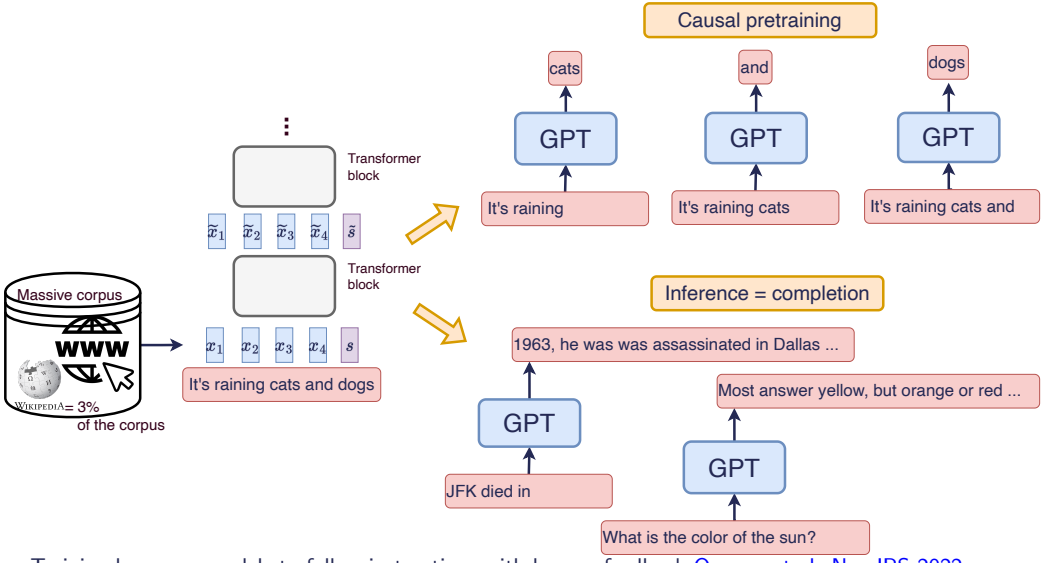
Disponibilité, possibilité de *fine-tuning*



Deep contextualized word representations, Peters et al. NAACL 2018

Les ingrédients de chatGPT

0. Transformer + données massives (GPT)

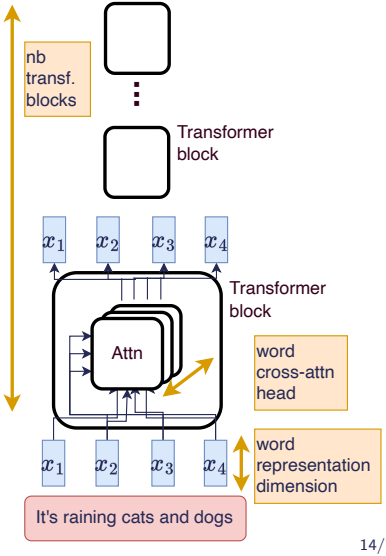


Training language models to follow instructions with human feedback Ouyang et al. NeurIPS 2022

Les ingrédients de chatGPT

1. Toujours plus! (GPT)

+ de mots en entrée [500 ⇒ 2k, 32k]
+ de dimensions (mots) [500-2k ⇒ 12k]
+ de têtes d'attention [12 ⇒ 96 (dim 128)]
+ de blocks/couches [5-12 ⇒ 96]

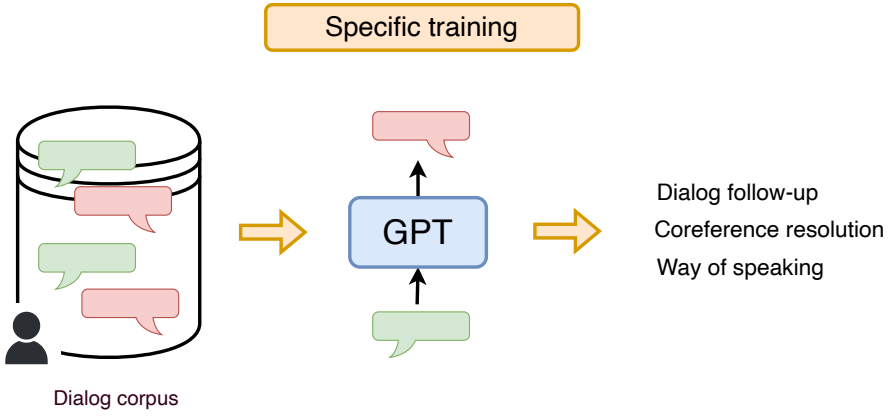


175 Milliards de paramètres... Ca fait quoi?

- small $1.75 \cdot 10^{11}$ ⇒ 300 Go + 100 Go (inférence) ≈ 400Go
- GPU NVidia A100 = 80Go de mémoire (=20k€)
- Coût pour (1) entrainement: 4.6 Millions d'€

Les ingrédients de chatGPT

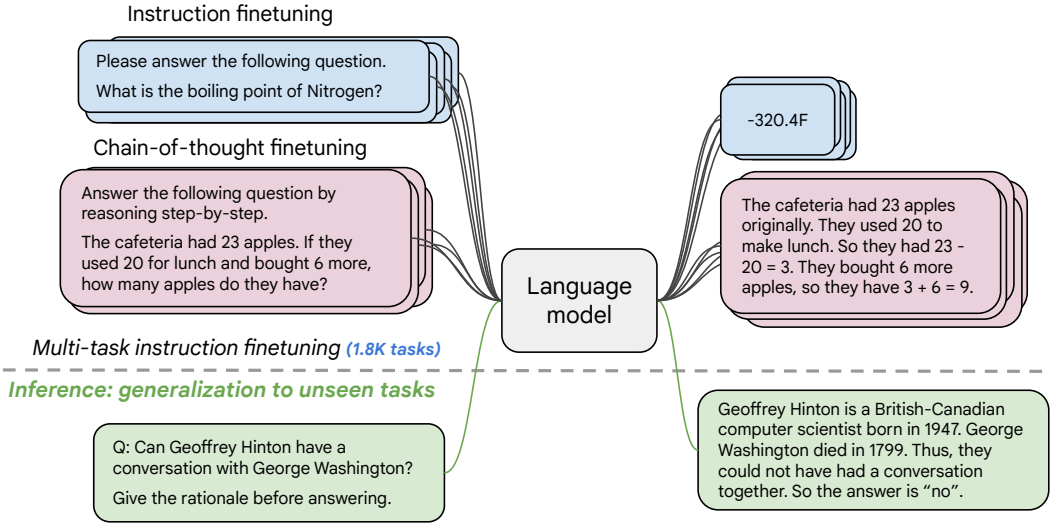
2. Suivi de dialogue



- Données **très propres** Données générées/validées par des humains

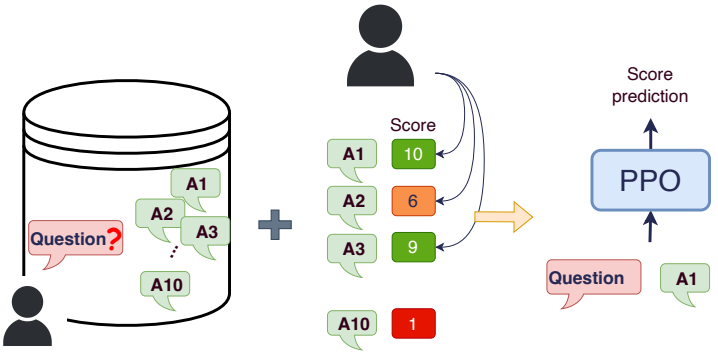
Les ingrédients de chatGPT

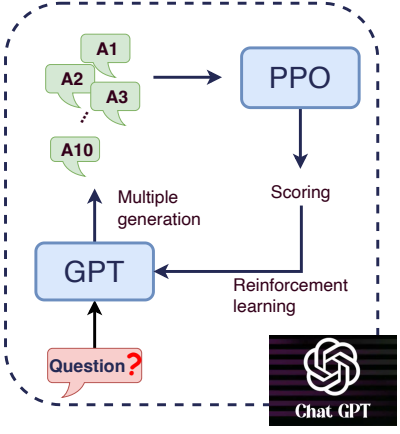
3. L'affinage sur différentes tâches de raisonnement ($\pm$) complexe



Les ingrédients de chatGPT

4. Suivi de dialogue & amélioration des réponses



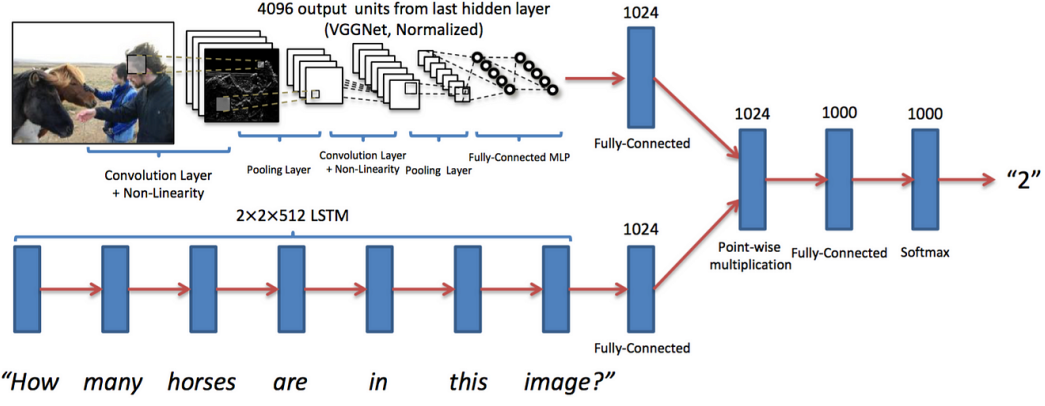


- BD faite par des humains
- Amélioration des réponses
- ... Aussi une manière d'éviter les sujets critiques

GPT4 & la multimodalité

Fusionner les informations issues du texte et de l'image.
Apprendre à exploiter les informations conjointement

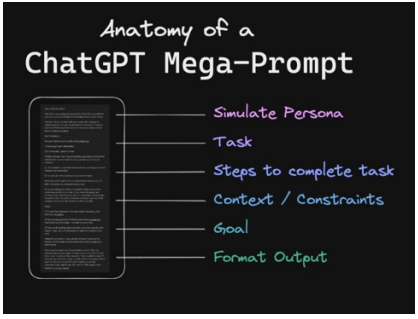
L'exemple du VQA: visual question answering



Rétro-propager l'erreur $\Rightarrow$ opti. représentations de mots + analyse image

Usage de chatGPT & Prompting

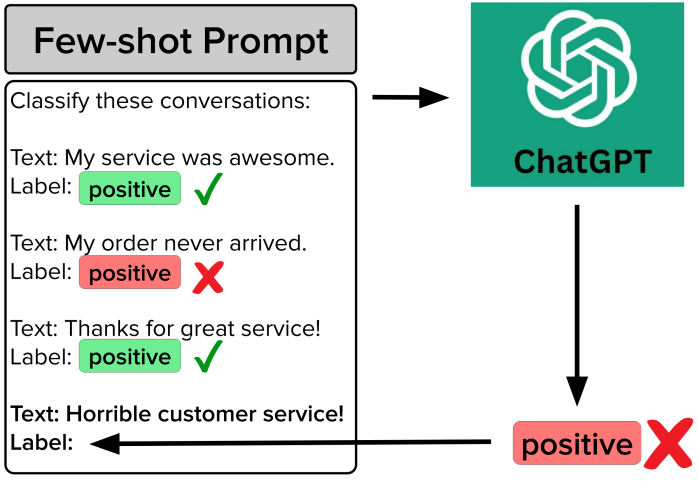
- Interroger chatGPT... Ca s'apprend! = *prompting*
  - Bien poser une question: *... en détails*, *... step by step*
  - Spécifier nb elts e.g. : *3 qualités pour ...*, *5 éléments pour...*
  - Poser un contexte : *cellule* pour un biologiste / assistant juridique
- Ne pas s'arrêter à la première question
  - Détailler des points particuliers
  - réorienter la recherche
- Reformulation
  - Explain like I'm 5, plus formel, à la manière d'un article scientifique, bro style, ...
  - Résumer, étendre
  - Ajouter des fautes (!)



https://chatgptprompts.guru/what-makes-a-good-chatgpt-prompt/

Vers du *few-shot learning*

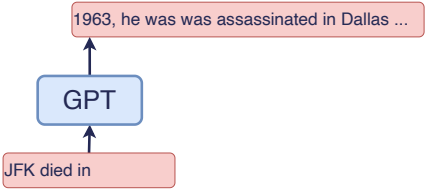
- Apprendre sans modifier le modèle = exemples dans le promt

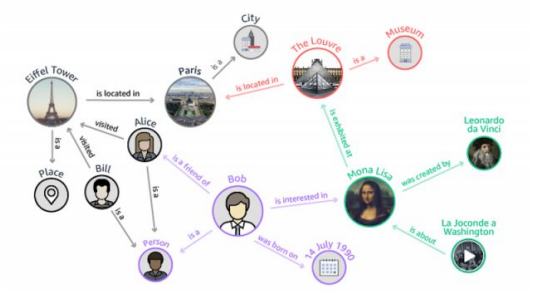


Enjeux et Limites

chatGPT et le rapport à la vérité

1. Vraisemblance = grammaire, accords, concordance des temps, enchaînements logiques...
   $\Rightarrow$ Connaissances répétées $\approx$ grammaire
2. Prédire le mot le plus **vraisemblable**...
   $\Rightarrow$ produit des **hallucinations**
3. Fonctionnement **hors-ligne**
4. chatGPT = loin des **graphes de connaissances**
5. Des réponses brillantes... Et des erreurs bêtes!
   + on ne sait pas prédire les erreurs

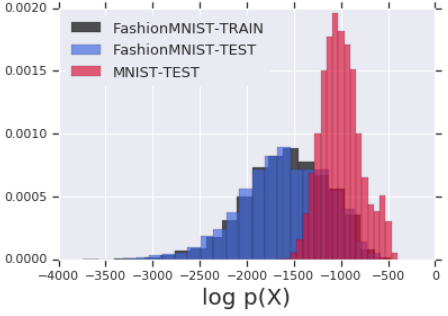




IA Génératives: comment évaluer les performances?

Le point critique aujourd'hui

- Comment s'évaluer par rapport à une vérité terrain?
- Comment évaluer la confiance du système / vraisemblance de la génération ?



Vraisemblance

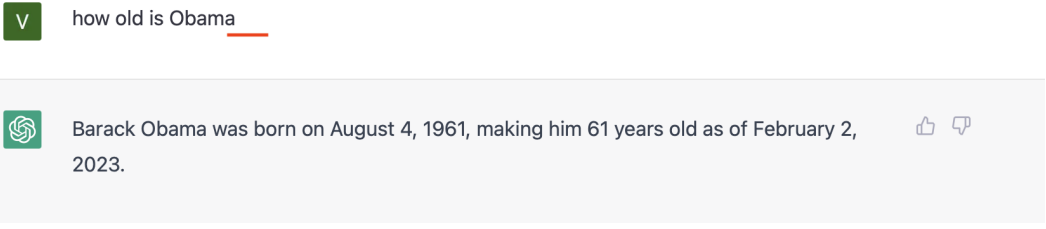
Train

Test

Stabilité/prédictibilité

- Difficile de borner un comportement
- Impossible de prédire les bonnes/mauvaises réponses

$\Rightarrow$ Peu/pas utilisé en jeux vidéo



Stabilité/prédictibilité

- Difficile de borner un comportement
- Impossible de prédire les bonnes/mauvaises réponses

⇒ Peu/pas utilisé en jeux vidéo

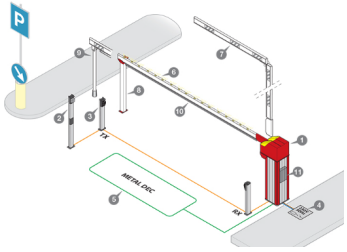
V how old is obama?

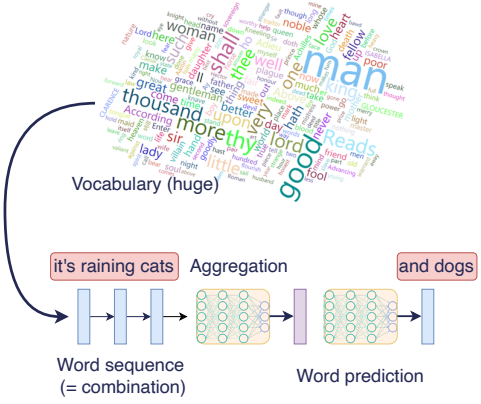
As of 2021, Barack Obama was born on August 4, 1961, so he is 60 years old.

V and today?

As a language model AI, I don't have real-time access to current dates. However, Barack

Stabilité, explicabilité... Et complexité





- Système *simple*
- Test exhaustif des entrées/sorties
- **prédictible** & **explicable**

- Grande dimension
- Combin. non-linéaires complexes
- **non-prédictible** & **non-explicable**

Stabilité, explicabilité... Et complexité

Interprétabilité *vs* Explication post'hoc

Réseaux de neurones = **non interprétable** (presque toujours)

trop de combinaisons pour anticiper

Réseaux de neurones = **explicable a posteriori** (presque toujours)

roles des entrées dans une décision sur un exemple

[Accident Uber, 2018]

- Système *simple*
- Test exhaustif des entrées/sorties
- **prédictible** & **explicable**

- Grande dimension
- Combin. non-linéaires complexes
- **non-prédictible** & **non-explicable**

Transparence

- Les poids du modèle (*open-weight*)... $\Rightarrow$ mais pas que les poids
- Les données d'entrainement (*BLOOM*) + distribution + instructions
- Techniques d'apprentissage
- Evaluation

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

Machine-learning & biais

Biais dans les données

- Biais dans les réponses

Oreilles pointues, moustaches, texture de poils
=
Chat

Homme blanc, +40ans, costume
=
Cadre supérieur

Le machine-learning est basé sur l'extraction de biais statistiques...
$\Rightarrow$ Lutter contre les biais = forcer l'algorithme à la main

Machine-learning & biais

Biais dans les données

- Biais dans les réponses

The nurse and the doctor ⇒ L'infirmière et le docteur

- Choix du genre
- Couleur de peau
- Posture
- ...

Le machine-learning est basé sur l'extraction de biais statistiques...

⇒ Lutter contre les biais = forcer l'algorithme à la main

Correction des biais & ligne éditoriale

Correction des biais:

- Sélection de données spécifiques, ré-équilibrage
- Censure de certaines informations
- Censure des résultats de l'algorithme

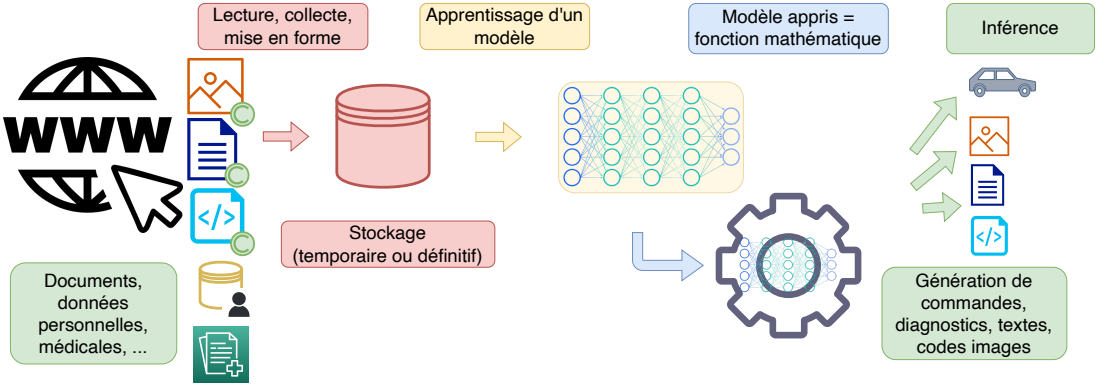
⇒ Travail éditorial...

Effectué par qui?

- Experts métiers / cahier des charges
- Ingénieurs, lors de la conception des algorithmes
- Groupe éthique, lors de la validation des résultats
- Groupe communication / réaction aux utilisateurs

⇒ Quelle légitimité? Quelle transparence? Quelle efficacité?

Risques/Questions juridiques



Droit d'auteur, droit des bases de données

Droit de collecte, droit de copie, consentement

Droit d'utiliser les données dans un algorithme

Modèle = émanation des données ?

Clearview.ai

Cambridge Analytica

Reproductions d'extraits non traçables

Régulation des usages

Responsabilité en cas d'erreur

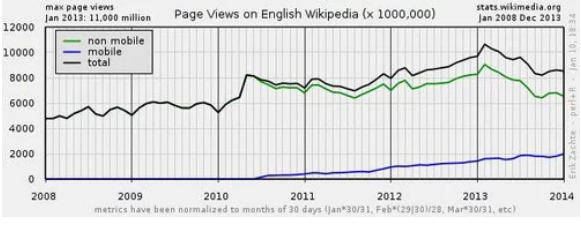
Questions économiques

Financement des sources d'information = publicité

- Publicité ⇔ **visites** des internautes
- Google knowledge graph (2012) ⇒ − de visites, − de revenu
- chatGPT = encodage des informations du web... ⇒ beaucoup moins de visites?

Google Knowledge Graph aurait causé une baisse du trafic de Wikipedia en 2013

Guillaume Belfiore
13 janvier 2014 à 12h14



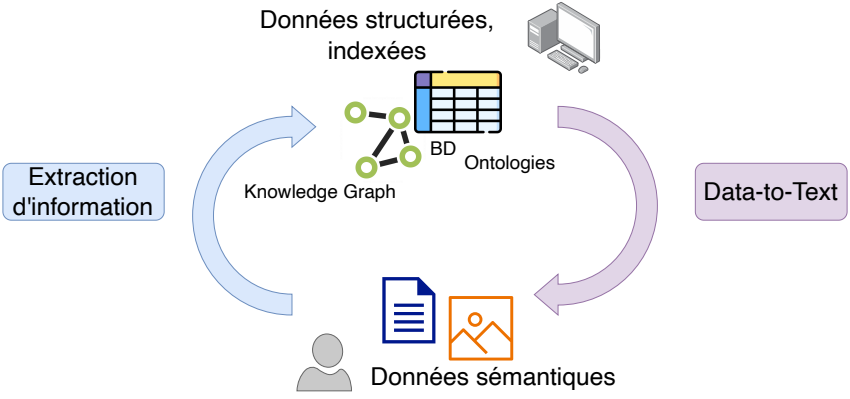
⇒ Quel **modèle économique pour les sources d'information** avec chatGPT?

Cycle des données: Extraction d'information Génération de texte

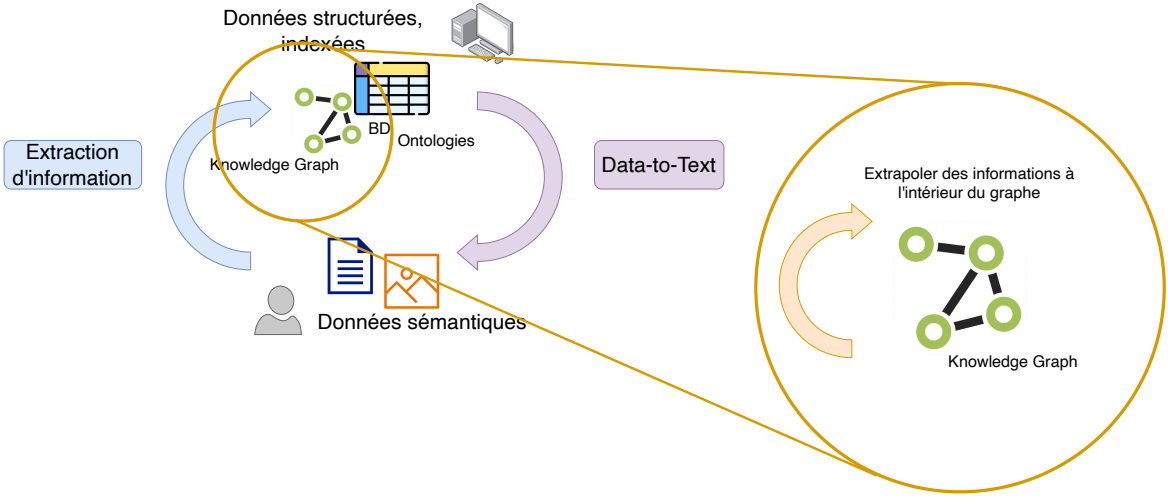
Cycle de l'information entre l'humain et la machine

Humain et machine: des modalités différentes, quel traducteur?

- Extraction d'information: actif depuis les années 80 [regex, pattern, etc...]
  $\Rightarrow$ révolutionné depuis 2012
- Génération de textes: idéal ancien, possibilités récentes [2014]



Enrichir les bases de connaissances



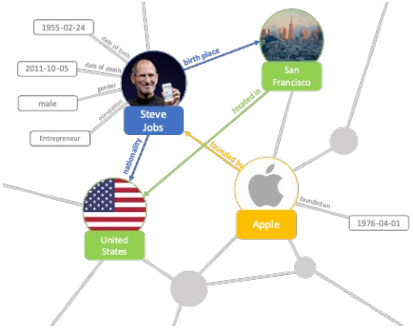
Enjeux autour des bases de connaissances

- Construire des bases de connaissances
- Raisonner: règles + inférence logique, ontologies, systèmes experts

Steve Jobs

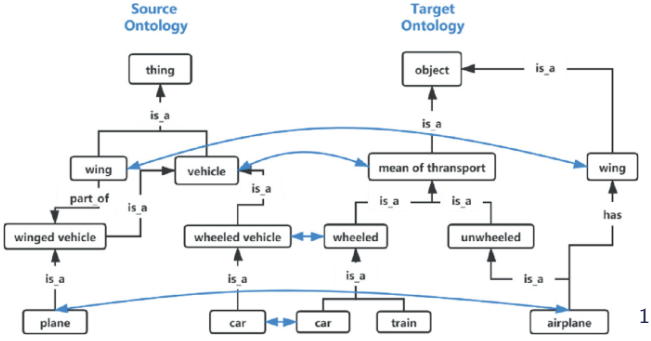
Jobs presenting the iPhone 4 in June 2010

Born	February 24, 1955 San Francisco, California, U.S.
Died	October 5, 2011 (aged 56) Palo Alto, California, U.S.
Resting place	Alta Mesa Memorial Park
Occupation	Entrepreneur · industrial designer · media proprietor · investor
Years active	1976–2011
Known for	Pioneer of the personal computer revolution with Steve Wozniak Co-creator of the Apple II, Macintosh, iPod, iPhone, iPad, and first Apple Stores
Title	Co-founder, chairman and CEO of Apple Inc. Co-founder, primary investor and chairman of Pixar Founder, chairman and CEO of NeXT
Board member of	The Walt Disney Company[1] Apple Inc.
Spouse(s)	Laurene Powell (m. 1991)
Partner(s)	Chrisann Brennan (1972–1977)



Enjeux autour des bases de connaissances

- Construire des bases de connaissances
- Raisonner: règles + inférence logique, ontologies, systèmes experts
- Connexions w/ Machine Learning
  - Alignement / fusion
  - Plongement / TransE



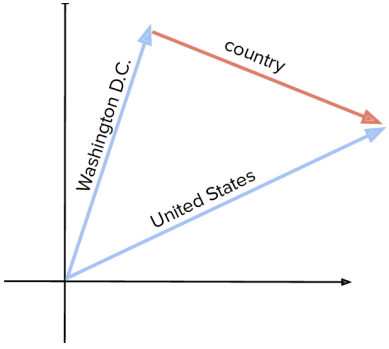
[1]

[1] Xiaojing Wu, Xingsi Xue, and Wenyu Hu (2021). "Argumentation Based Ontology Alignment Extraction". In: Advanced Machine Learning Technologies and Applications. ISBN: 978-3-030-69717-4

Enjeux autour des bases de connaissances

- Construire des bases de connaissances
- Raisonner: règles + inférence logique, ontologies, systèmes experts

- Connexions w/ Machine Learning
  - Alignement / fusion
  - Plongement / TransE

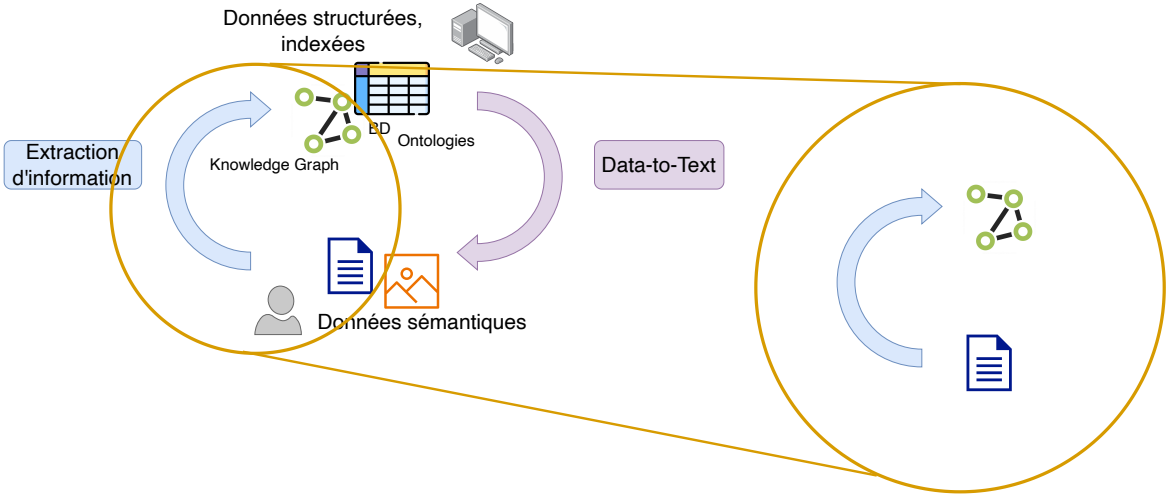


[1]

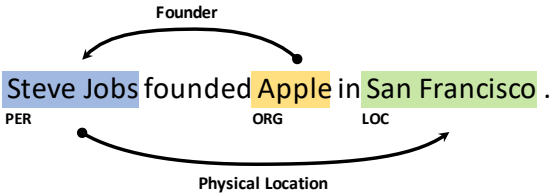
[1] Antoine Bordes et al. (2013). "Translating embeddings for modeling multi-relational data". In: NeurIPS

Extraction d'Information

Du texte au connaissances structurées



Challenges autour de l'extraction d'information



- Segmenter les entités
- Identifier et/ou typer les entités
- Identifier + classer les liens

- Défi de la segmentation: e.g. New York Times
- Polysémie
- Fautes d'orthographe ⇒ Morphologie + sémantique + contexte

Boston

Washington

Philadelphia

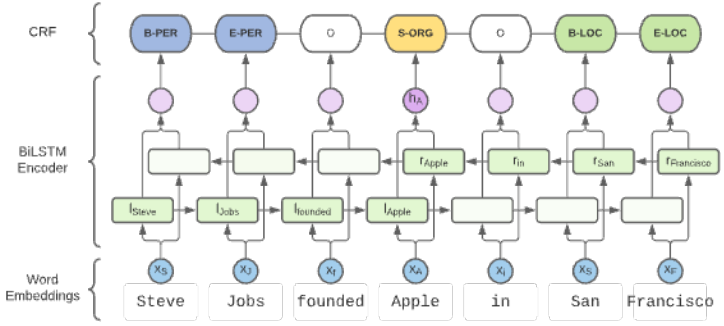
Extraction des entités nommées

Steve	Jobs	founded	Apple	in	San	Francisco	.
B-PER	**E-PER**	O	**S-ORG**	O	**B-LOC**	**E-LOC**	O

IOBES : **O = Other (not in an entity)**
B = Beginning **I = Inside** **E = End** **S = Single**

2

Révolutions successives: représentation des mots & contextualisation



- **Pretrained word embeddings**
  (Huang 2015) SENNA
- **Character-level word embeddings**
  (Lample 2016) SENNA + char-BiLSTM
- **Contextualized embeddings**
  (Peters 2018) ELMo
  (Akbik 2018) Flair
  (Devlin 2019) BERT

[2] Bruno Taillé (2022). "Contextualization and Generalization in Entity and Relation Extraction". PhD thesis. Sorbonne Université

Extraction des entités nommées

ELMo (Peters 2018)

- **char-CNN** word representation **(ELMo[0])**
- **BiLSTM** LM at a **word** level
- Weighted sum fusion (learned weights)

Flair (Akbik 2018)

- **BiLSTM** LM at a **character** level
- Word represented with the concatenation of its ends

BERT (Devlin 2019)

- **Transformer** LM at a **subword** level (WordPiece)
- Masked LM and Next Sentence Prediction
- **$BERT_{LARGE}$ feature-based = frozen LM**

CoNLL03 Test Set (F1)		
$BERT_{LARGE}$	(Devlin 2019)	92.8
ELMo	(Peters 2018)	92.2
Flair	(Akbik 2018)	92.0*
TagLM (SENNA + LM)	(Peters 2017)	91.9
SENNA + char BiLSTM	(Lample 2016)	90.9
SENNA	(Huang 2015)	88.8

(Peters 2018) Deep contextualized word representations, NAACL-HLT 2018
(Akbik 2018) Contextual String Embeddings for Sequence Labeling, COLING 2018
(Devlin 2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL-HLT 2019

Une tâche résolue?

Superposition lexicale: apprentissage *vs* test

Proportion of mentions in test set are seen during training.

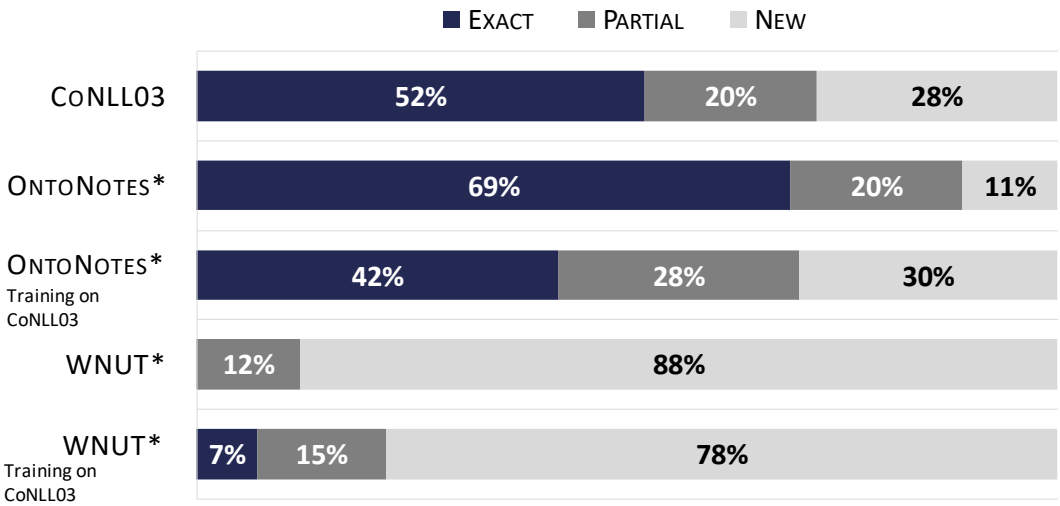
3 types of mentions :

Exact match	Mention seen with the same type
Partial match	At least one non stop-word seen in a mention of same type
New	All non stop-words are new

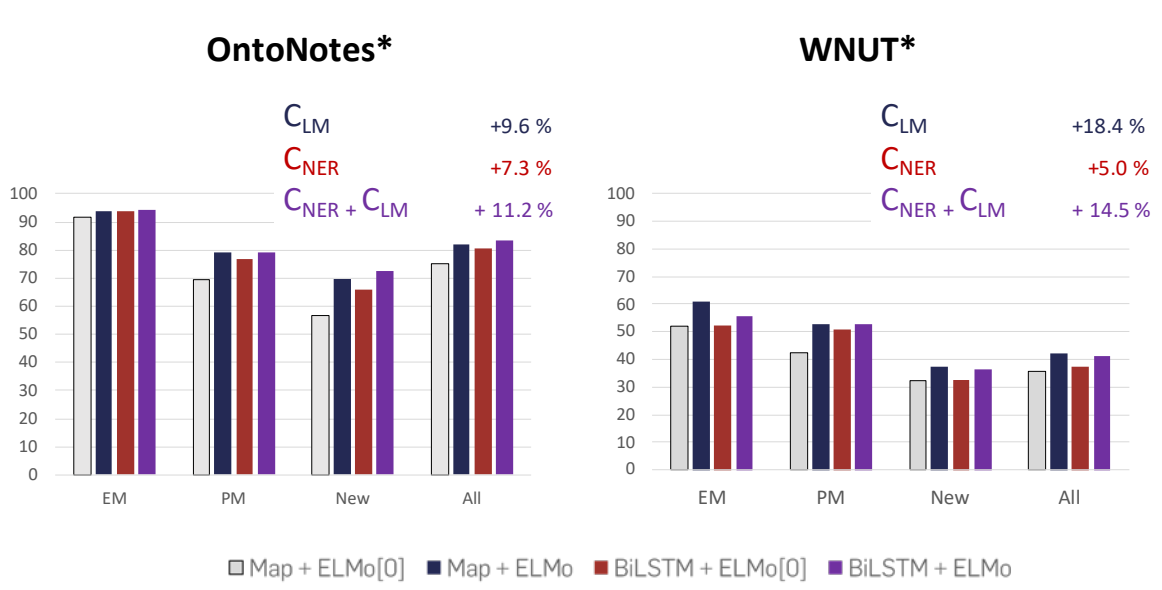
Train :	Georges Washington	(PER)
	Barack Obama	(PER)
Test :	**Donald Trump**	**(PER)**
	Barack Obama	**(PER)**
	Georges Bush	**(PER)**
	Washington DC.	**(LOC)**
	Obama	**(PER)**

(Augenstein 2017) Generalisation in named entity recognition: A quantitative analysis, CSL 2017
(Moosavi 2017) Lexical Features in Coreference Resolution: To be Used With Caution, ACL 2017

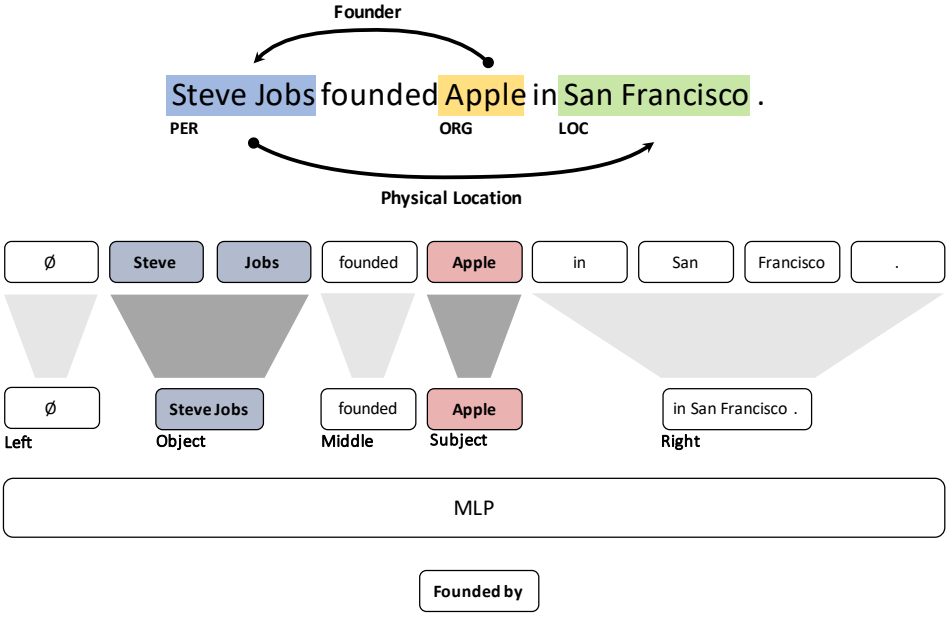
Superposition lexicale: apprentissage *vs* test



Séparation des performances: les résultats



Extraction de relation: pipeline & piecewise pooling



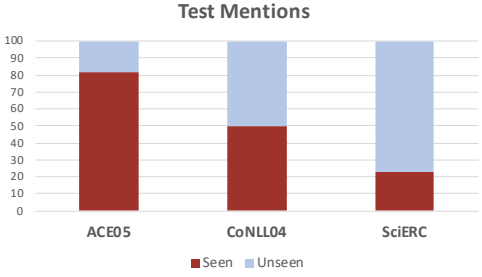
Superposition des ensembles d'apprentissage & test

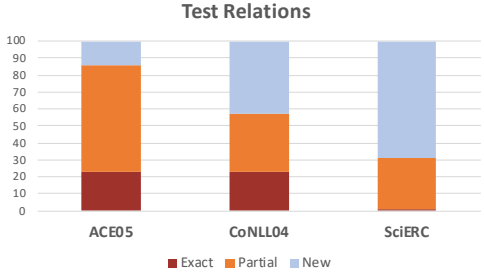
NER
(Augenstein 2017, Taillé 2020)

- **Seen** — Exact Match with the same type
- **Unseen**

RE

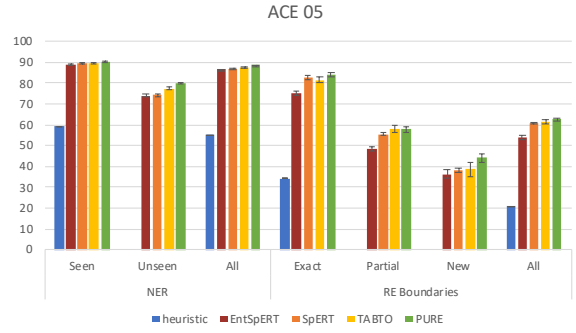
- **Exact Match** — Triple **(head, predicate, tail)** exactly seen during training
- **Partial Match** — **(head, predicate, ...)** or **(..., predicate, tail)** seen during training
- **New** — Otherwise

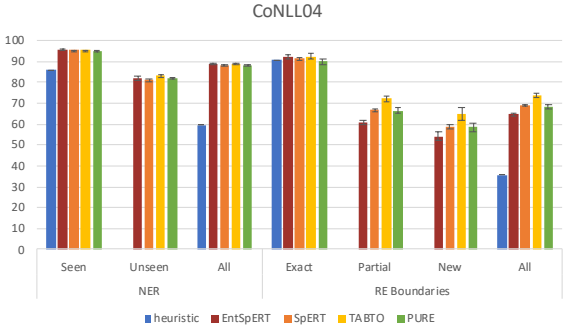




Relation Extraction *vs* End-to-end Relation Extraction

Superposition des ensembles d'apprentissage & test





Relation Extraction *vs* End-to-end Relation Extraction

De nombreux défis autour de l'extraction d'information

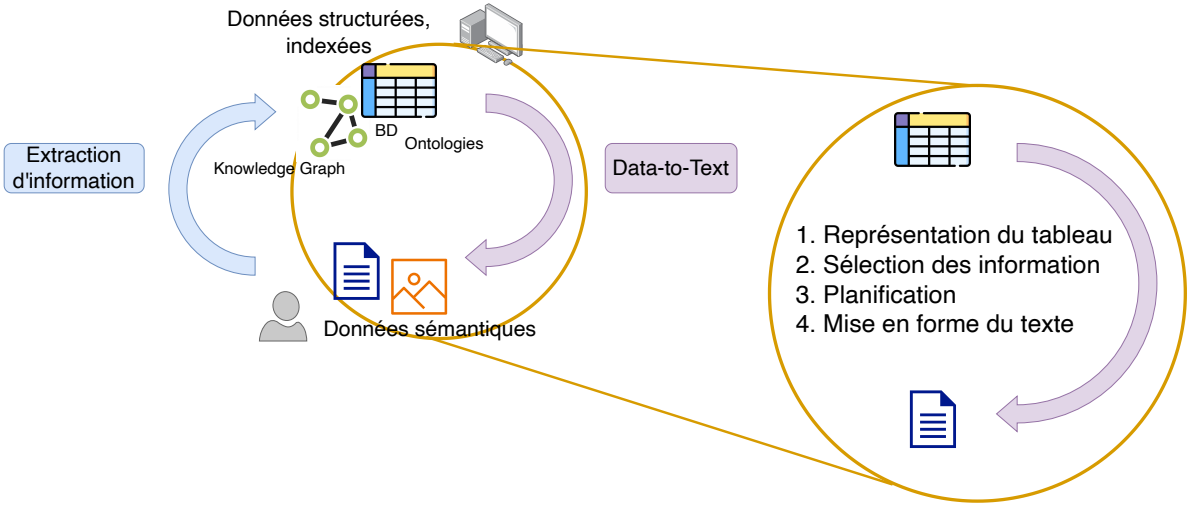
- Extraction des entités & des relations [Thèse de B. Taillé]
- Etiquetage distant / auto-supervision des modèles [Thèse de M. Sahraoui]
- Reconnaissance d'entité dynamique (dont la classe change) [Thèse de T. Luiggi]
- Contextualisation des entités / désambiguisation [Thèse de T. Luiggi/T. Herserant]
- Exploitation des IA Générative pour la tâche / mesure de performances [Thèse de T. Luiggi/T. Herserant]

$\Rightarrow$ Une problématique très ouverte

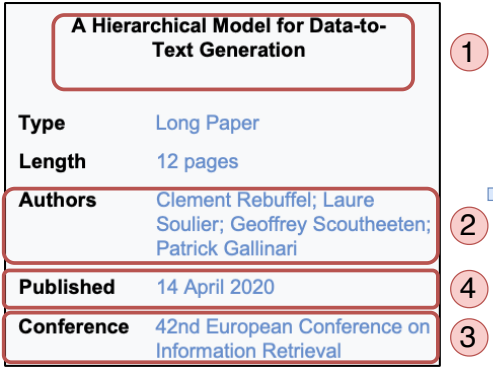
▸ Extensions

Data-to-text

Un processus complexe en plusieurs étapes



Exemple (1)



A Hierarchical Model for Data-to-Text Generation (Rebuffel et al.) will be published at ECIR 2020

- Content Selection
- Macro-planning
- Surface realisation
- Sentence aggregation
- Data abstraction/interpretation

Exemple (2)

	Fund	Benchmark	Excess
OFFICIAL PERFORMANCE (net)	3,39%	5,34%	-1,96%
OFFICIAL PERFORMANCE (gross)	3,48%	5,34%	-1,86%
INTERNAL PERFORMANCE	3,10%	5,34%	-2,24%

	Allocation	Selection	FX rates	CGT Prov	Residual
MANAGEMENT EFFECTS	-0,81%	-1,42%	0,46%	-0,01%	-0,08%

Reference Benchmark : MSCI China A, NR
Report Currency : EUR

SECURITY LEVEL - BIGGEST CONTRIBUTORS TO EXCESS RETURN

Company Name	Grouping	Perf	Var(W)	Effect
China Pacific Insu	*Financials*	13,43	3,15	0,25
Shanxi Lu'an Envir	*Energy*	14,82	2,12	0,21
Anhui Conch Cement	*Materials*	13,41	2,56	0,20
China Minsheng Ban	*Financials*	10,61	3,58	0,18
Jizheng Energy Res	*Energy*	13,61	2,20	0,18
Jiangsu Yueda Inv	*Industrials*	11,78	2,24	0,14
Poly Real Estate G	*Financials*	9,14	3,50	0,13
Xinjiang Ba Yi Iro	*Materials*	16,69	0,89	0,10
China Nonferrous	*Materials*	13,84	0,96	0,08
China Baoan Group	*Industrials*	17,34	0,61	0,07

SECURITY LEVEL - BIGGEST DETRACTORS FROM EXCESS RETURN

Company Name	Grouping	Perf	Var(W)	Effect
Nari Tech Dev Co	*Industrials*	-6,48	3,26	-0,39
Shandong Denghai S	*Consumer Staples*	-8,32	1,76	-0,26
Zte Corp	*Information Technology*	-9,38	1,66	-0,25
Mesnac Co Ltd	*Industrials*	-9,17	1,54	-0,22
Yunnan Baiyao Grp	*Health Care*	-6,82	1,70	-0,22
Jiangsu Aoyang Tec	*Materials*	-14,77	0,49	-0,22
Fujian Septwolves	*Consumer Discretionary*	-7,86	1,46	-0,20
Tianjin Tasly Phar	*Health Care*	-9,11	1,05	-0,17
Xi An Aero-Engine	*Industrials*	-3,10	1,83	-0,16
Ping An Insurance	*Financials*	13,27	-1,96	-0,15

SECTOR LEVEL - BEST ALLOCATION DECISIONS

Grouping	Var(W)	Segm. Perf	Effect
Health Care	-2,24	-5,53	0,25
Consumer Discretionary	-4,65	3,12	0,11
Energy	2,51	9,23	0,10
Industrials	0,19	5,23	0,01
Utilities	0,29	1,67	-0,01

SECTOR LEVEL - BEST SELECTION DECISIONS

Grouping	PF Perf	BM Perf	Effect
Financials	9,58	8,69	0,22
Information Technology	-1,93	-2,63	0,07
Telecommunication Services	-0,89	-2,01	0,01

Materials	7,37	11,15	-0,35
Consumer Discretionary	-2,19	3,12	-0,28

In January (30/12/2011 to 20/1/2012), Flexifund Equity China A rose in value by **3.48%** compared to a gain of **5.34%** for its index in **Euro** terms. Both **asset allocation** and **stock selection** detracted from relative performance, as the market focused on oversold or cyclical themes, due to better global risk appetite and more positive economic news.

From a sector allocation perspective, […]

Exemple (3)

Toronto Raptors (4-2)

Player	MIN	ORTG	USG%	PTS	FG	3PT	FT	OREB	DREB	TO	AST	BLK	STL	PF
Pascal Siakam F	46	132.0	20.4	26	10-17	3-6	3-4	2	8	2	3	1	1	2
Kawhi Leonard F	41	115.7	23.5	22	7-16	1-5	7-8	1	5	2	3	1	2	4
Marc Gasol C	27	98.8	12.8	3	0-5	0-2	3-4	3	6	1	4	0	0	4
Kyle Lowry G	42	137.3	23.2	26	9-16	4-7	4-6	2	5	3	10	0	2	5
Danny Green G	18	-	2.5	0	0-0	0-0	0-0	0	1	1	3	0	1	1
Fred VanVleet G	34	132.5	22.7	22	6-14	5-11	5-5	1	1	1	0	0	1	1
Serge Ibaka C	22	118.7	28.5	15	7-12	0-1	1-2	2	1	1	2	0	0	4
Norman Powell G	11	-	12.4	0	0-2	0-1	0-0	0	1	1	0	0	0	2
Game Total	**240**	-	-	**114**	**39-82**	**13-33**	**23-29**	**11**	**28**	**12**	**25**	**2**	**8**	**23**

Golden State Warriors (2-4)

Player	MIN	ORTG	USG%	PTS	FG	3PT	FT	OREB	DREB	TO	AST	BLK	STL	PF
Draymond Green F	44	92.9	19.0	11	5-10	1-4	0-2	4	15	8	13	2	3	4
Andre Iguodala F	32	117.1	25.0	22	9-15	3-6	1-5	0	2	1	2	1	0	3
Kevon Looney C	27	109.4	13.2	6	3-7	0-0	0-0	2	1	1	4	1	1	2
Stephen Curry G	42	113.3	23.7	21	6-17	3-11	6-6	1	2	3	7	1	2	3
Klay Thompson G	32	100.6	25.0	30	8-12	4-6	10-10	1	4	2	0	0	2	3
DeMarcus Cousins C	19	117.3	27.6	12	4-9	0-1	4-7							
Shaun Livingston G	16	92.0	16.7	6	3-5	0-0	0-0							
Quinn Cook G	18	78.2	10.3	2	1-3	0-2	0-0							
Alfonzo McKinnie F	10	-	4.4	0	0-1	0-1	0-0							
Andrew Bogut C	3	-	-	0	0-1	0-0	0-0							
James Jerebko F	2	-	-	0	0-0	0-0	0-0							
Game Total	**240**	-	-	**110**	**39-80**	**11-31**	**21-30**							

The **Toronto Raptors** defeated the host **Golden State** Warrior, **114-110**, in Game 6 of the NBA Finals at ORACLE Arena on Thursday. [...]

The **Raptors** (4-2) were lead by **Kyle Lowery**, as he accrued **26 points, seven rebounds, 10 assists and three steals**. [...]

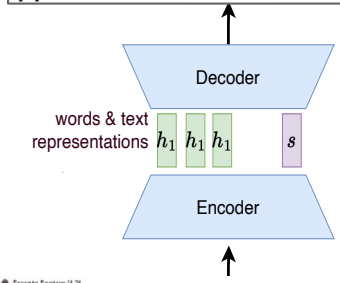
[...]

Comment aborder les données structurées ?

The **Toronto Raptors** defeated the host **Golden State** Warrior, **114-110**, in Game 6 of the NBA Finals at ORACLE Arena on Thursday. [...]

The **Raptors** **(4-2)** were lead by **Kyle Lowery,** as he accrued **26 points, seven rebounds, 10 assists and three steals**. [...]

[...]



Toronto Raptors (4-2)

Player	MIN	OREB	DREB	REB	AST	STL	BLK	TO	PF	PTS
Pascal Siakam F	46	2	8	10	3	1	1	2	2	20
Kawhi Leonard F	41	1	5	6	3	1	2	2	4	26
Marc Gasol C	37	3	6	9	4	3	0	3	4	14
Kyle Lowry G	42	2	5	7	10	3	0	3	5	22
Danny Green G	18	0	1	1	3	0	1	1	1	7
Fred VanVleet G	34	1	1	2	3	1	0	1	1	11
Serge Ibaka C	22	2	1	3	2	0	1	3	4	12
Norman Powell G	11	0	0	0	0	0	0	0	2	2
Team Total	240	11	28	39	12	5	2	15	23	114

- R Resume
- T Table

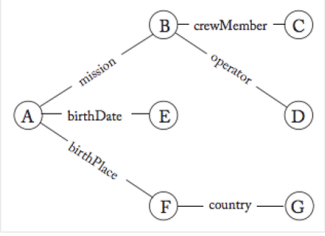
$$P(R \mid T, \theta) = \prod_{i=1}^{\ell} P\left(y_i \mid y_{<i}, T, \theta\right)$$

Comment encoder la table?

Linéarisation de table & pre-training

Country	PTS
Germany	2
Argentina	0

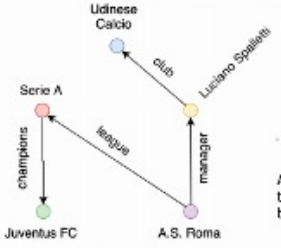
[(0, Country, Germany), (0, PTS, 2), (1, Country, Argentina), (1, PTS, 0)]



Input: *A (birthPlace F (country G)) (birthDate E) (mission B (operator D) (crewMember C))*

(Narayan and Gardent et al. 2020) Deep Learning Approaches to Text Production

Linéarisation de table & pre-training



<S> Serie A <P> champions <O> Juventus F.C. <S> Luciano Spalletti <P> club <O> Udinese Calcio <S> A.S. Roma <P> manager <O> Luciano Spalletti <S> A.S. Roma <P> league <O> Serie A

AS Roma play in the Serie A league where Juventus FC are the champions. Their manager is Luciano Spalletti who has been associated with Udinese Calcio.

Domain	train
Inform	arrive_by : 11:51
Request	num_people

train inform arriveby = 11:51 | train request people = ?

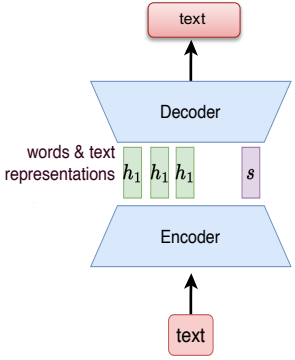
The closest arrival time i can give you is 11:51 , is that ok ? And how many tickets would you like ?

Table Title: Cristhian Stuani
Section Title: International goals

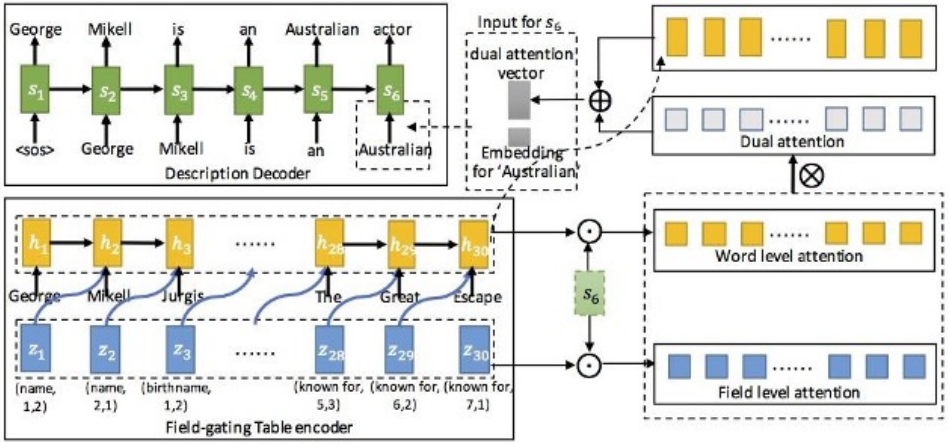
No.	Date	Venue	Opponent	Result
2	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	5-0

<page_title> Cristhian Stuani </page_title> <section_title> International goals </section_title> <table> <cell> 2. <col_header> No. </col_header> </cell> <cell> 13 November 2013 <col_header> Date </col_header> </cell> <cell> Amman International Stadium, Amman, Jordan <col_header> Venue </col_header> </cell> <cell> Jordan <col_header> Opponent </col_header> </cell> <cell> 5-0 <col_header> Result </col_header> </cell> </table>

On 13 November 2013 Cristhian Stuani netted the second in a 5–0 win in Jordan.



Amélioration: encoder & sélectionner les valeurs du tableau

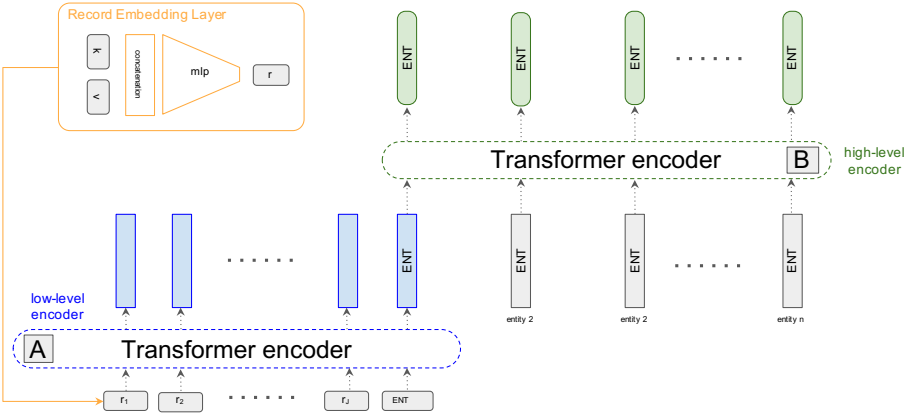


- Scores d'attention sur les mots et les champs de la table

$\Rightarrow$ Apprendre à encoder + sélectionner les informations

(Liu et al. 2018) Table-to-text Generation by Structure-aware Seq2seq Learning

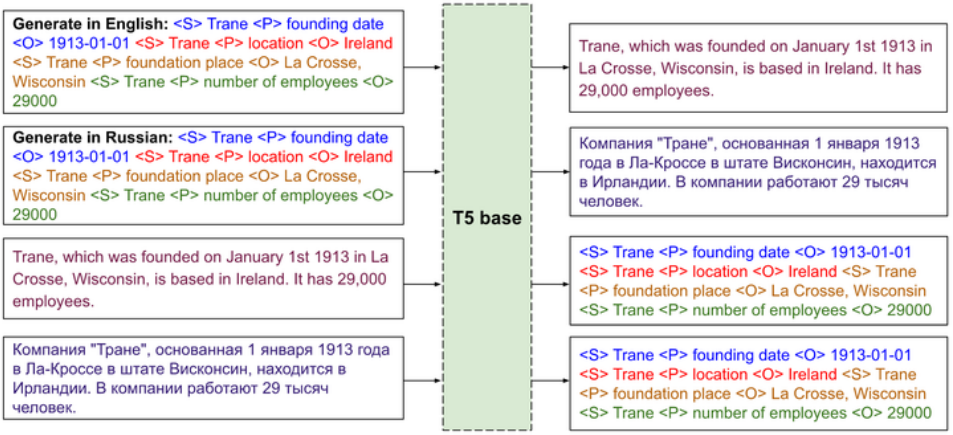
Amélioration: encoder & sélectionner hiérarchiquement



- Encodage hiérarchique sur Rotowire (stats des joueurs match basket):
  - Encodage d'une case du tableau - ref. colonne
  - Encodage d'une ligne du tableau (token [ENT]) - ensemble du joueur
  - Encodage du tableau

(Rebuffel et al. 2021) A Hierarchical Model for Data-to-Text Generation, ECIR 2021

Et au bout du compte?



- T5: un modèle *tout en un* dédié aux traduction des données structurées
- Passage D2T et T2D possible avec le même modèle

(Agarwal et al. 2020) Machine Translation Aided Bilingual Data-to-Text generation and Semantic Parsing

Grandes pathologies de la génération de texte

Attribute	Value
Birthplace	*Utah, America*
Position	*forward (soccer player)*

Omission

A soccer player, who plays as a forward.

- Contenu attendu mais manquant dans le texte généré

Hallucination

A Utah forward, from the national team.

- Texte généré contient du texte divergent de la table
- Textes de références souvent divergents

- Des connaissances issues du LLM interfèrent
- Sur-apprentissage de pattern de sélection de la base d'apprentissage

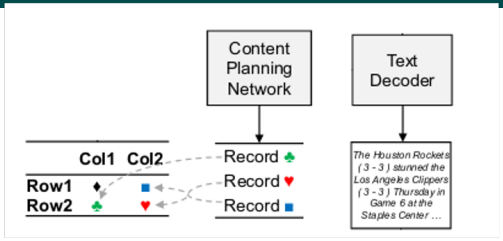
(Liu et al. 2019) Example from Towards Comprehensive Description Generation from Factual Attribute-value Tables

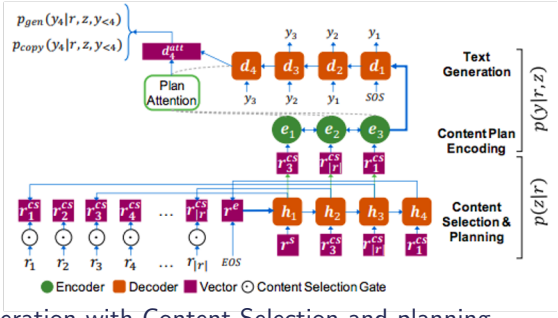
Améliorer la génération par la planification

- Guidage = lutte contre les hallucinations

1. Génération d'un plan (séquence ordonnées de clés-valeurs)
2. Génération du texte final

Variante : génération séquentielle d'un élément du plan et de la phrase associée





(Puduppully et al. 2018) Data-to-Text Generation with Content Selection and planning.
(Puduppully et al. 2022) Data-to-text Generation with Variational Sequential Planning

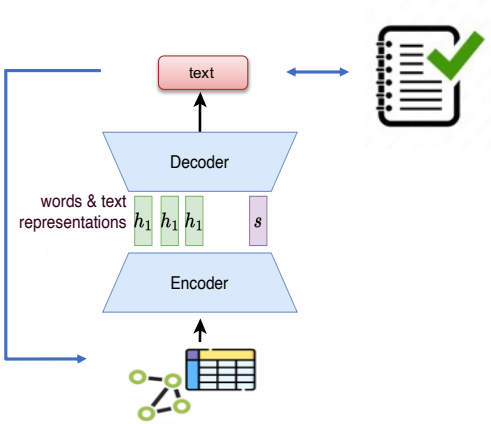
Optimiser directement la génération

1. Trouver la bonne métrique: PARENT
   métrique d'appariement entre le texte généré et les données structurées (n-grams, co-occurrences)
2. Optimiser la métrique non dérivable

⇒ Apprentissage par renforcement

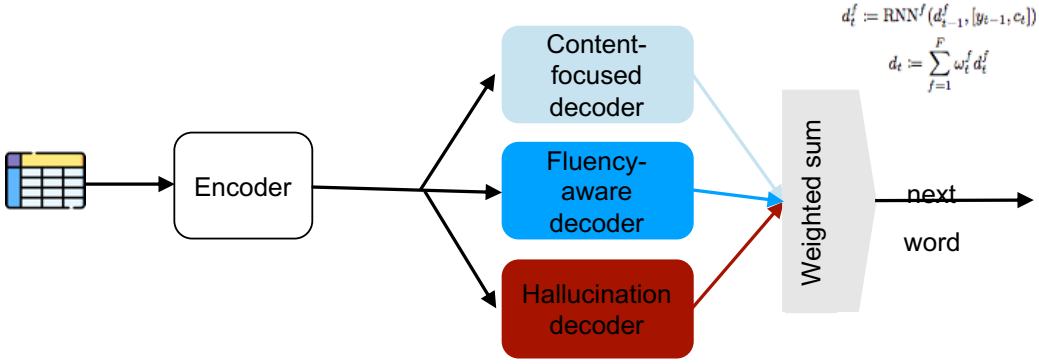
RL loss based on PARENT — Cross-Entropy loss

$$\mathcal{L} := \gamma \mathcal{L}_{rl} + (1-\gamma)\mathcal{L}_{ml}$$



(Rebuffel et al. 2020) PARENTing via Model-Agnostic Reinforcement Learning to Correct

Distinguer les hallucinations lors du décodage



Architecture multi-branches

- **Supervision** très fine des phrases générées
- Séparation des **différents générateurs** (RNN) + Scores
- **Balance** lors de la génération

(Rebuffel et al. 2022) Controlling hallucinations at word level in data-to-text generation, DMKD 2022

Apprentissage contrastif

Lutter contre les hallucinations en *Question-Answering*

1. Modèle conditionné à la table et la question
2. Modèle simple (conditionné par la question seule)

Donnée structurée (entrée)	Texte attendu	Texte bruité
<H> AMC_Matador <R> bodyStyle <T> Coupé	The **AMC Matador's body style** is **Coupé**	The **Alc Matador** is a **Spanish bourgeois coupe**.
name[Clowns] eatType[coffee shop] food[Fast food] customer rating[high] area[riverside] near[Clare Hall]	**Clowns** is a **coffee shop** which offers **fast food** and has **high customer ratings**, and may be found near **Clare Hall** in the **riverside** area	**Cats** and **coffee shop**, and the **fast food** place, **Clowns**, is located near **Clare Hall**. It is in the **riverside** area. It has a **high customer rating**.

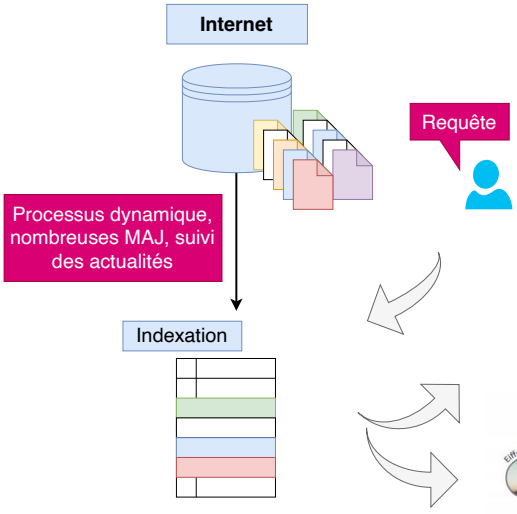
$\Rightarrow$ Apprendre à éliminer les hallucunations = cout contrastif entre générations

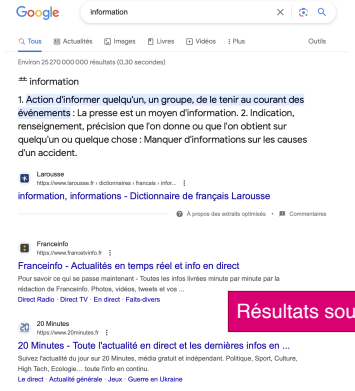
$\Rightarrow$ PPO/DPO pour l'apprentissage

(LeBronnec et al. 2024) Rédaction en cours :)

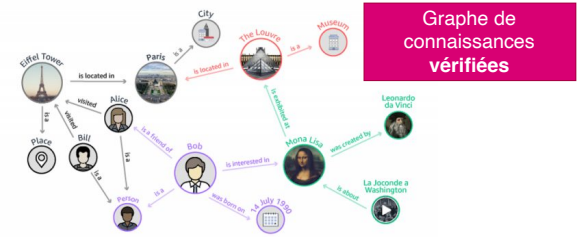
Accès à l'information et Modèle de Langue

Usage en accès à l'information





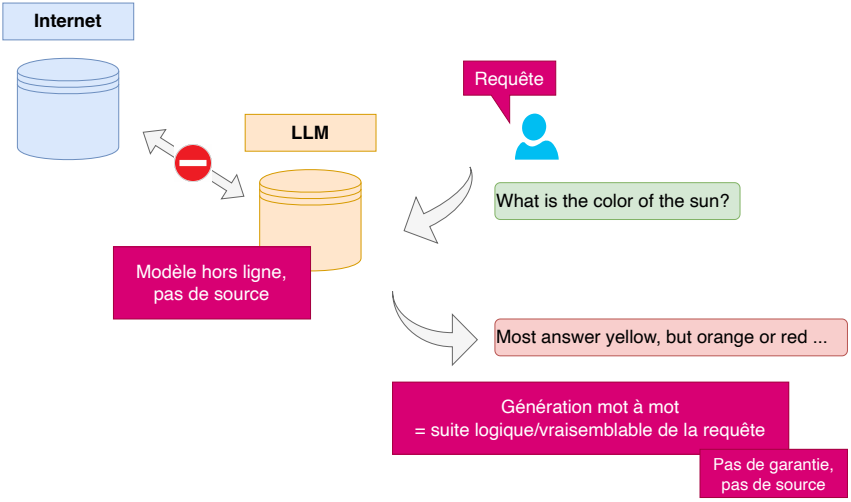
Résultats sourcés



Graphe de connaissances **vérifiées**

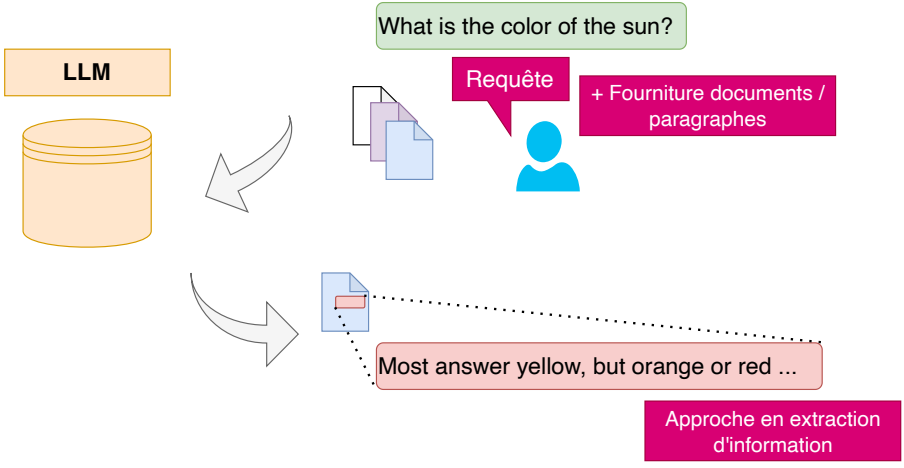
Usage en accès à l'information

- Demander des informations à chatGPT... Un usage étonnant !



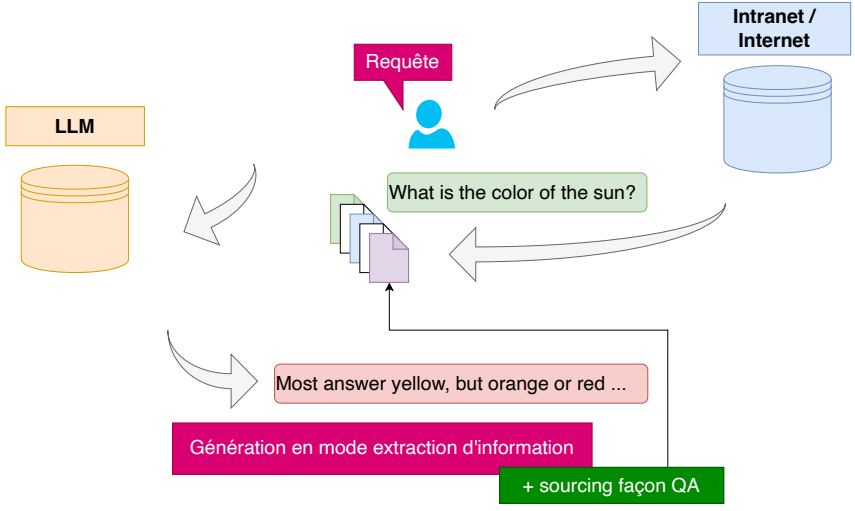
- LLM limité en connaissances
- Risque d'hallucination à la génération

Usage en accès à l'information



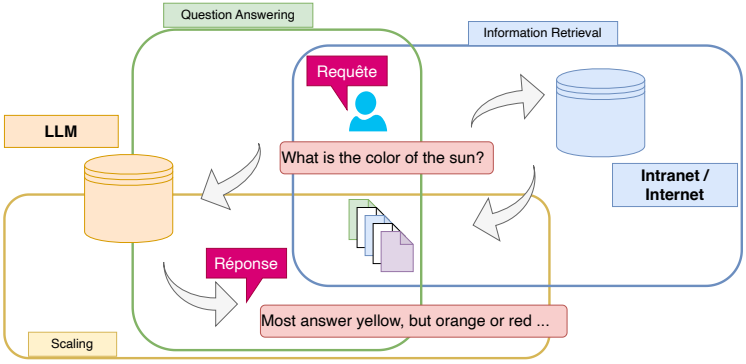
- Requête web + analyse, résumé automatique, reformulation, compte-rendus de réunion...
- Limite (actuelle) sur la taille des entrées (2k puis 32k puis 100k tokens)

Usage en accès à l'information



- RAG: *Retrieval Augmented Generation*
- Limite (actuelle) sur la taille des entrées (2k puis 32k puis 100k tokens)

L'état de l'art en RAG



Retrieval-Augmented Generation (RAG) [1]

Improve performance on knowledge intensive task (question answering)

Retrieval-Augmented Language Model (REALM) [2]

Integrate retrieval augmented into the pre-training

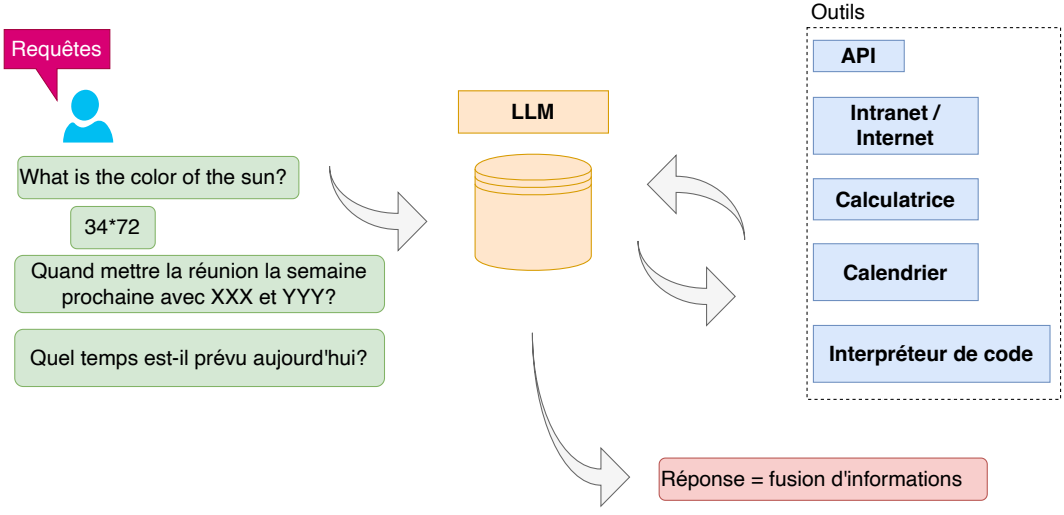
Retrieval-Enhanced Transformer (RETRO) [3]

Scale generation to large number of retrieved documents

[1] Guu et al (2020), REALM: Retrieval-Augmented Language Model Pre-Training
[2] Lewis et al (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
[3] Borgeaud et al (2022) Improving Language Models by Retrieving from Trillions of Tokens

Multiplier les outils: le LLM / couteau Suisse

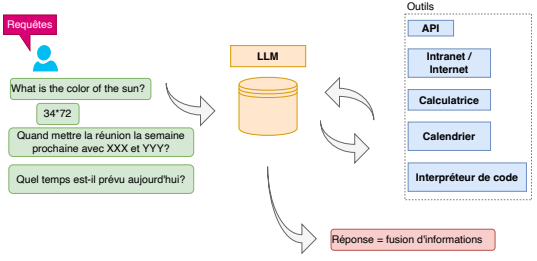
- Apprendre au LLM à appeler (*balise*) des outils externes



(Schick et al. 2023), Toolformer: Language Models Can Teach Themselves to Use Tools

Multiplier les outils: le LLM / couteau Suisse

- Apprendre au LLM à appeler (*balise*) des outils externes



The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

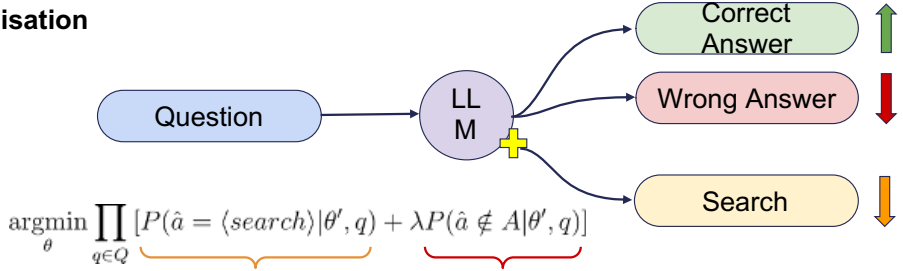
The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Optimiser le cout des outils

Objectif : Apprendre à générer le token <SEARCH> lorsque cela est nécessaire

Formalisation

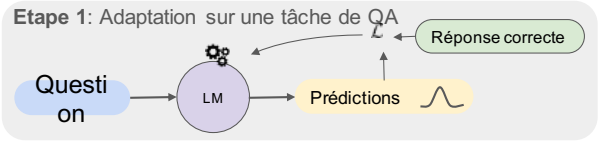


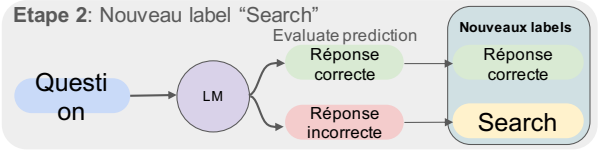
- Toolsformer appelle le moteur de recherche dans 99% des cas
- Peut-on faire la balance avec les connaissances du LLM?

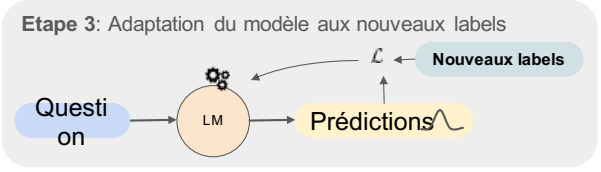
Optimiser le cout des outils

Apprendre une fonction de filtrage qui :

- Laisse inchangé les Correct Answer
- Masque les Wrong Answer avec Search

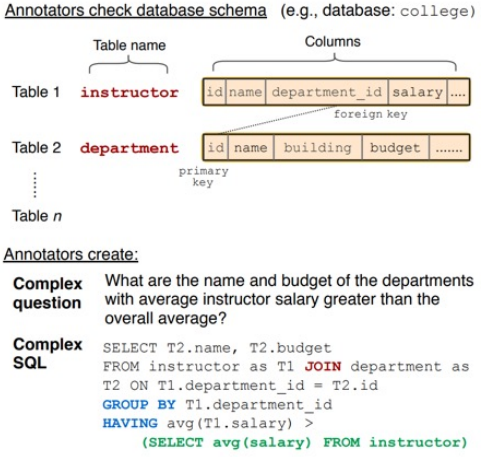






(Erbacher et al. 2023) Navigating Uncertainty: Optimizing API Dependency for Hallucination Reduction in Closed-Book QA, ECIR 2023

Le SQL: un outil comme les autres?



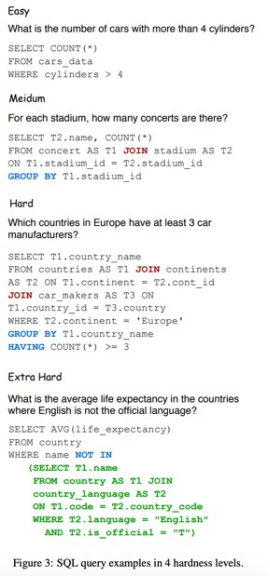


Figure 3: SQL query examples in 4 hardness levels.

- TableQA: schema + question $\Rightarrow$ SQL
- Comprendre ce qui est facile ou dur

(Yu et al. 2018) Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic

Le SQL: un outil comme les autres?

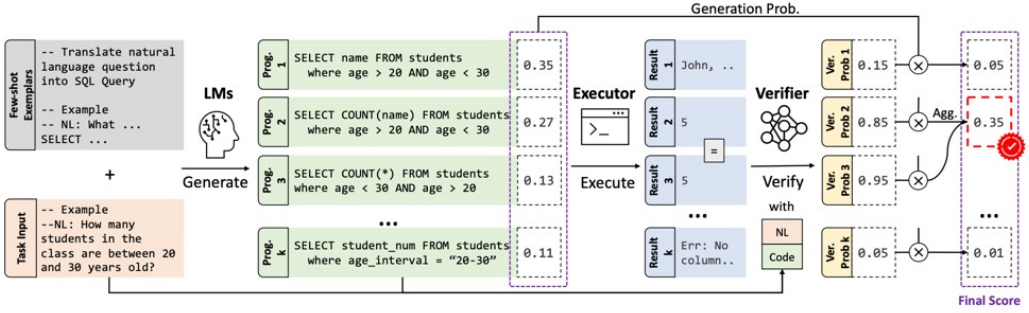
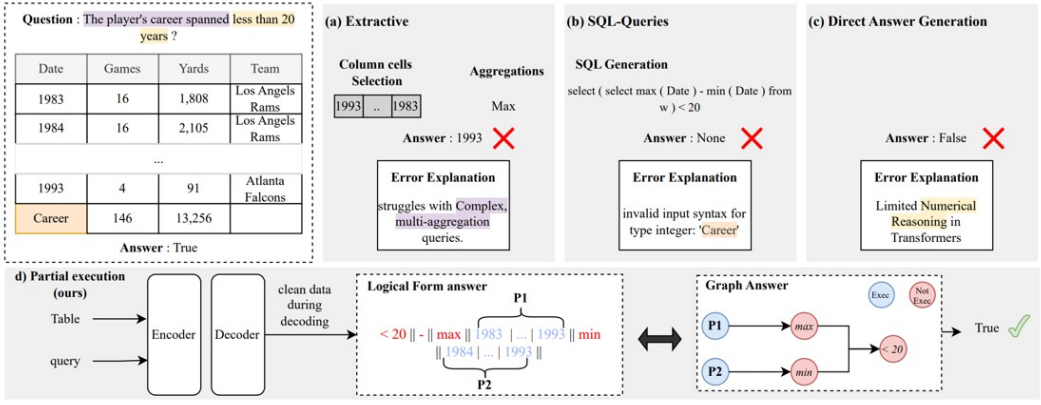


Figure 1: The illustration of LEVER using text-to-SQL as an example. It consists of three steps: 1) *Generation*: sample programs from code LLMs based on the task input and few-shot exemplars; 2) *Execution*: obtain the execution results with program executors; 3) *Verification*: using a learned verifier to output the probability of the program being correct based on the NL, program and execution results.

- Prédire les bonnes et les mauvaises réponses
- Plus de feedback pour mieux apprendre

(Ni et al. 2023), LEVER: Learning to Verify Language-to-Code Generation with Execution

Le SQL: un outil comme les autres?

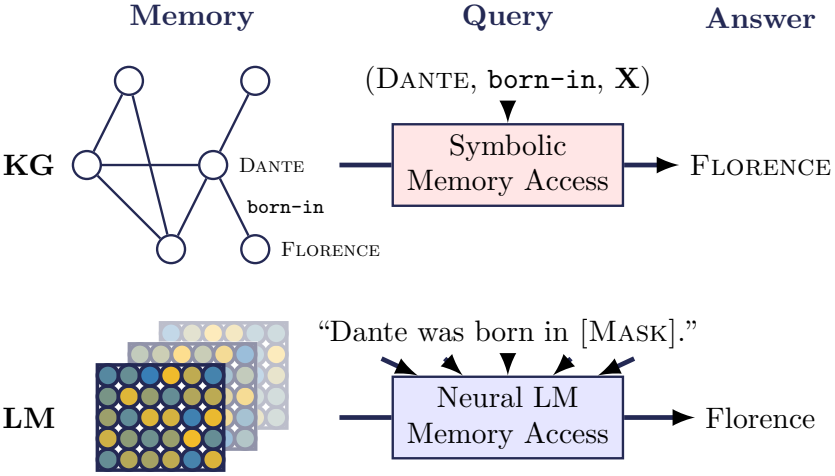


- Apprendre à raisonner numériquement à partir d'une base étiquetée en SQL
- Le LLM apprend à évaluer les requêtes SQL

(Mouravieff et al. 2024), Training Table Question Answering via SQL Query Decomposition

Conclusion

Sous quelle forme stocker les connaissances?

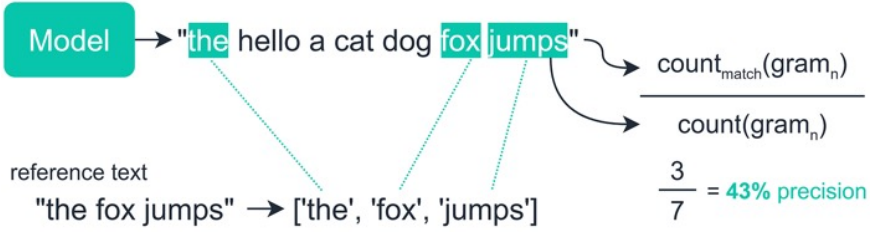


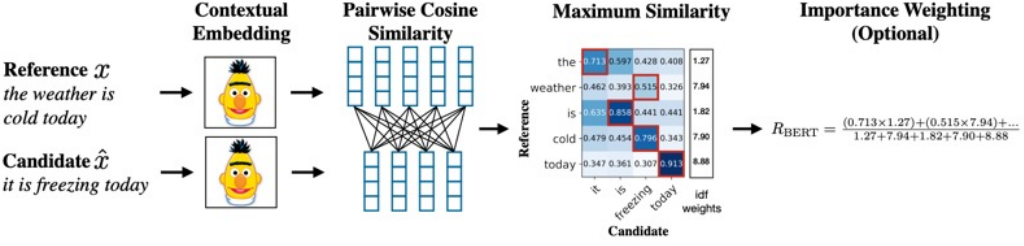
[2]

- Exhaustivité?
- Fiabilité?

[2] Fabio Petroni et al. (2019). "Language Models as Knowledge Bases?" In: EMNLP. Association for Computational Linguistics

Comment évaluer les modèles de langue?





Comment évaluer la qualité d'un texte ou d'une image générée ?

Conclusion et perspective

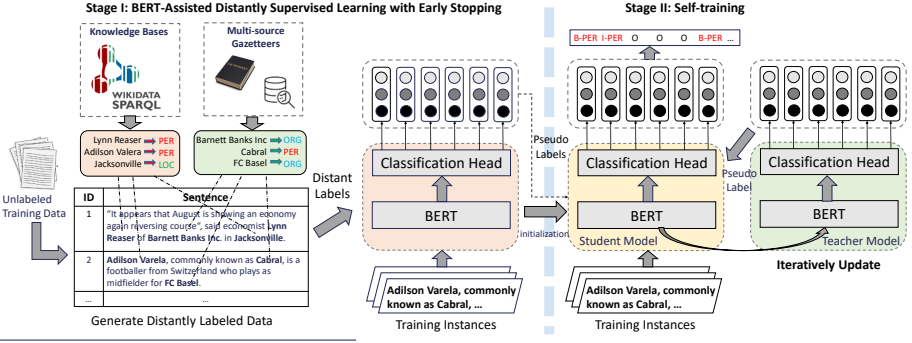
- LLM + Instruction = le début d'un mouvement
  - Objet de recherche dépassé par les usages
- Des technologies **chères** (mais un coût en baisse)
  - Ressources disponibles: Jean Zay
- Des limites critiques:
  - Evaluation
  - Controle / garantie sur la génération

Extraction d'Information: les nombreux défis

Résoudre l'annotation: auto-supervision (ex-distillation)

Processus NER standard

1. Liste d'entités
2. Etiquetage automatique du corpus — `regex`
3. Inférence sur le test
4. Ré-apprentissage — Teacher-student



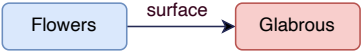
[3]

[3] Chen Liang et al. (2020). "Bond: Bert-assisted open-domain named entity recognition with distant supervision". In: ACM SIGKDD

Application à l'analyse des descriptions de flores

Extraction d'Information ⇒ Clé-valeur

<ORGAN> Flowers </ORGAN> 4-merous. Calyx aestivation valvate, campanulate, 2-3.6mm long, abaxially <DESC-SURFACE> glabrous </DESC-SURFACE>



solitary flowers; bracts 4–8, chartaceous, ovate or transverse-elliptic, 0.4–1.6 × 0.4–1.5 mm, marginally ciliolate with eglandular hairs, apically obtuse, obtuse and cuspidate, or acute, abaxially glabrous; pedicel 1–1.2 mm long, reduced and hidden by overlapping bracts, glabrate with eglandular hairs; differentiated apical bracteoles 2, distinct, chartaceous, partially enveloping calyx lobes, covering 50–67% of calyx, ovate, 1.5–2(–2.5) × 1.6–3 mm, marginally ciliolate or ciliate with eglandular hairs, apically obtuse and cuspidate or less often acuminate, the surface smooth, abaxially and adaxially glabrous. **Flowers** 4-merous. Calyx aestivation valvate, campanulate, (2–)2.4–3.3 mm long; tube slightly angled, 0.8–1.3 mm long,

TABLE I
STATISTICS ON THE DATASET : CLASSES, NUMBER OR DISTINCT WORDS IN EACH CLASS AND NUMBER OF OCCURRENCES IN THE CORPUS.

Set	Class	Occurrences	Number of words
$\mathcal{Y}_0$	Flower	22890	23
	Fruit	4968	10
	Habit	1920	3
	Leaf	4364	5
	Part-of	23849	25
	Stem-root	3296	7
$\mathcal{Y}_1$	Color	18342	15
	Disposition	8405	21
	Form	24816	64
	Position	10936	13
	Surface-texture	18325	23

4

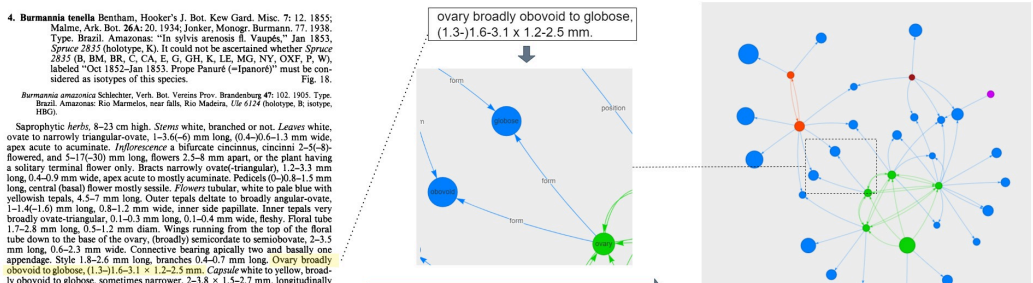
[4] Maya Sahraoui et al. (2022). "NEARSIDE: Structured kNowledge Extraction frAmework from Species DEscriptions". In: Biodiversity Information Science and Standards

Application à l'analyse des descriptions de flores

Models	Precision	Recall	Score F1
Baseline	100/93.83	75.74/70.82	86.19/79.26
Baseline w/ lm	100/95.15	85.28/80.82	92.05/86.54
Baseline w/self-train	100/94.42	84.29/80.15	91.47/86.22

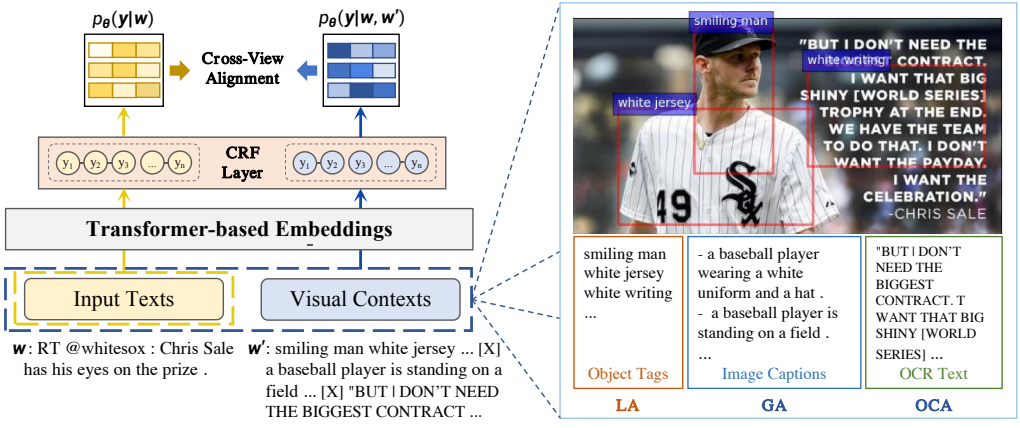
MODEL'S ABILITY TO DETECT AND CLASSIFY NEW ENTITIES, OUT OF THE TRAIN SET'S DISTRIBUTION. (DETECTION/CLASSIFICATION SCORES)

Models	Precision	Recall	Score F1
Baseline	100/92.33	64.78/54.52	78.62/62.76
Baseline w/ lm	100/89.88	69.21/57.73	81.80/65.17
Baseline w/self-train	100/90.76	68.95/57.82	81.62/64.90



Perspective: extension vers la multimodalité

⇒ Retrouver les **entités dans les images** à partir d'approche texte/image



[4]

[4] Xinyu Wang et al. (2022). "ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition". In: NAACL

Perspective: extension vers la multimodalité

⇒ Retrouver les **entités dans les images** à partir d'approche texte/image

solitary flowers; bracts 4–8, chartaceous, ovate or transverse-elliptic, 0.4–1.6 × 0.4–1.5 mm, marginally ciliolate with eglandular hairs, apically obtuse, obtuse and cuspidate, or acute, abaxially glabrous; pedicel 1–1.2 mm long, reduced and hidden by overlapping bracts, glabrate with eglandular hairs; differentiated apical bracteoles 2, distinct, chartaceous, partially enveloping calyx lobes, covering 50–67% of calyx, ovate, 1.5–2(–2.5) × 1.6–3 mm, marginally ciliolate or ciliate with eglandular hairs, apically obtuse and cuspidate or less often acuminate, the surface smooth, abaxially and adaxially glabrous. **Flowers** 4-merous. Calyx aestivation valvate, campanulate, (2–)2.4–3.3 mm long; tube slightly angled, 0.8–1.3 mm long,

⇒ Construire des systèmes pédagogiques pour l'identification de taxons

Dynamic NER

Cas extrême où les entités changent de type tout le temps!

Exemple: détecter les joueurs de NBA... Avec le résultat du match:

victoire/défaite

A trio of 20 - point - plus efforts and a 17 - rebound night helped hand the Cavs a surprising home loss , their first defeat of the season overall . Dennis Schroder ' s season - high 28 points led the way , while Kent Bazemore put together a stellar 25 - point tally while often going up against LeBron James ' typically stingy defense . Dwight Howard dominated down low with 17 boards , 15 of them on the defensive glass . Atlanta managed a strong 51 percent success rate from the field , helping to key the victory . Kyrie Irving posted 29 points , which came on a season - high 27 shot attempts . Kevin Love ' s 24 - point , 12 - rebound double - double was next , while LeBron James posted 23 points . Poor shooting was Cleveland ' s undoing , as they posted a 37 percent success rate from the field , and 26 percent on 42 shot attempts from beyond the arc .

Same entity
Different context
Different Label

LeBron James and Kyrie Irving stepped up for a second straight night in Kevin Love ' s absence , combining for 60 points on 23 - of - 41 shooting . Irving added a career - high 13 assists , six rebounds and a steal , while James posted nine rebounds and six assists . Richard Jefferson supplied 10 points in Love ' s stead , and Tristan Thompson hauled in 15 rebounds . A pair of 10 - point efforts from Channing Frye and Iman Shumpert paced the second unit . Giannis Antetokounmpo ' s 28 points led Milwaukee , and Jabari Parker was right behind him with 27 points , as the duo tried to keep pace with Cleveland ' s Big Two . However , John Henson , Tony Snell , and Matthew Dellavedova , the remaining members of the first unit , could only combine for nine points between them . Malcolm Brodgon supplied 11 points off the bench as the only other double - digit scorer .

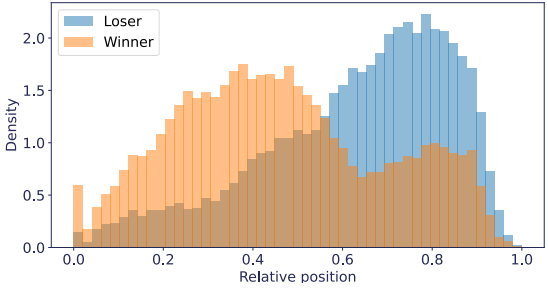
4

[4] Tristan Luiggi et al. (2023). "Dynamic Named Entity Recognition". In: ACM SAC

Discussion D-NER

- Proposition de nouvelle ressource

Models	Set	RotoWire DNET	RotoWire DNER	RotoWire Entity
BERT-Linear	Seen	0.81	0.66	0.86
	Seen/Unseen	0.81	0.65	0.85
	Unseen	0.80	0.63	0.81
BERT-CLS	Seen	0.81	**0.67**	0.88
	Seen/Unseen	0.81	**0.68**	0.87
	Unseen	0.80	**0.67**	0.85
BERT-CRF	Seen	-	0.67	**0.90**
	Seen/Unseen	-	0.67	**0.88**
	Unseen	-	0.66	**0.87**
BERT-CLS-CRF	Seen	-	0.61	0.82
	Seen/Unseen	-	0.61	0.81
	Unseen	-	0.60	0.79

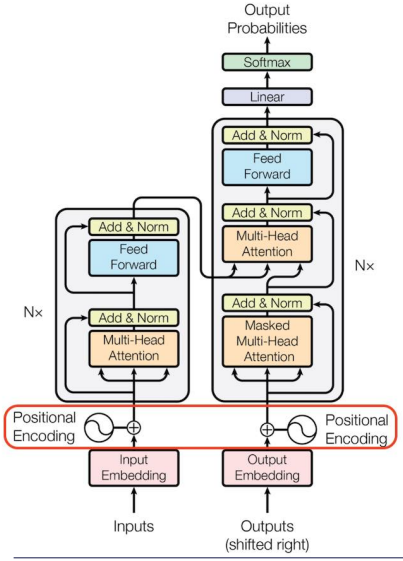


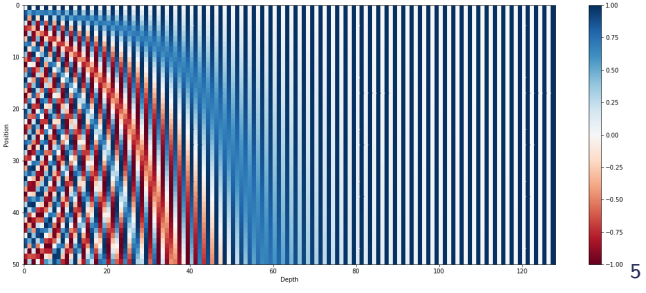
- Analyse de l'intérêt de la couche CRF
- Distinction in-domain / ood

$\Rightarrow$ Des perspectives vers l'encodage de la position des mots

Détection des entités dans un document structuré

Encodage de la position dans les documents:

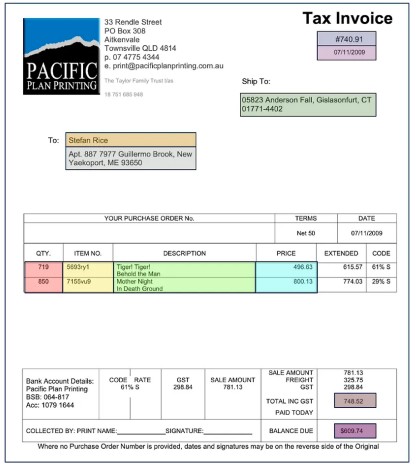




[5]

[5] A. Kazemnejad (2019). Transformer Architecture: The Positional Encoding.

Détection des entités dans un document structuré



- Invoice Number
- Invoice Date
- Shipping Address
- Customer Name
- Billing Address
- Quantity
- SKU
- Description
- Unit Price
- Total
- Balance Due

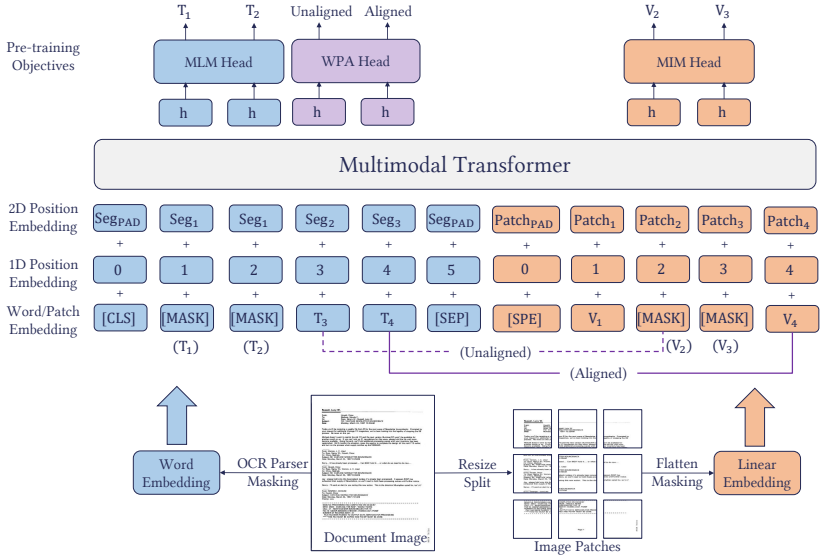
- Texte
- Image
- Coordonnées des mots

Puiser dans les modalités pour améliorer les performances

5

[5] Yiheng Xu et al. (2020). "Layoutlm: Pre-training of text and layout for document image understanding". In: ACM SIGKDD

Détection des entités dans un document structuré



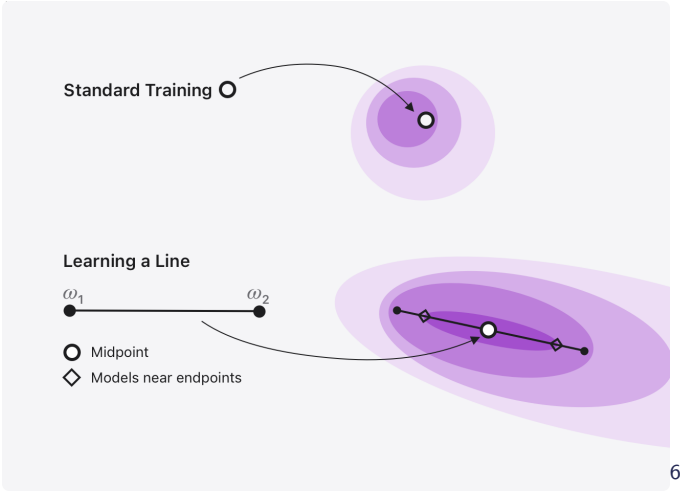
5

⇒ A quel moment souhaite-t-on mélanger les modalités?

[5] Yupan Huang et al. (2022). "LayoutLMv3: Pre-training for Document AI with Unified Text and Image

Optimisation robuste pour la généralisation

Optimisation de sous-espaces



6

Création de *régions homogènes* dans l'espace de représentation

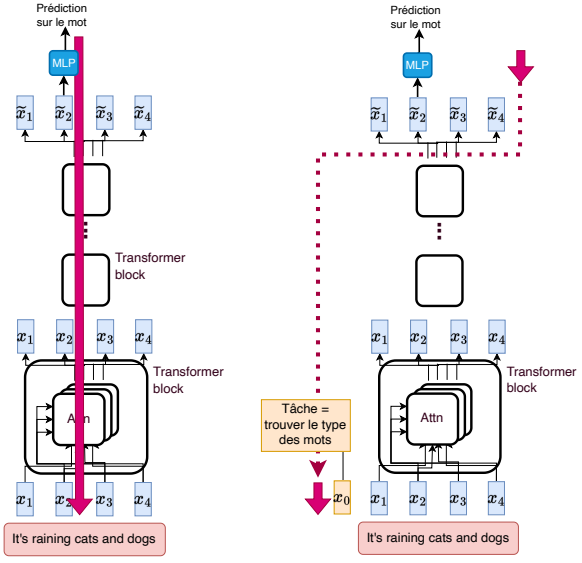
⇒ Améliorer l'espace de représentation

[6] Mitchell Wortsman et al. (2021). Learning Neural Network Subspaces.

Prefix-tuning & optimisation

- Impossible de maintenir plusieurs versions des paramètres d'un LLM
- Possible de travailler sur des approches parcimonieuses

⇒ Amélioration dans diverses tâches GLUE... Mais pas encore en NER[7]



[7] Louis Falissard, Vincent Guigue, and Laure Soulier (2023). "Improving generalization in large language models by learning prefix subspaces". In: EMNLP

Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?

Comment analyser la phrase suivante?

`Azawad reprend les armes`

Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?

En allant chercher du contexte sur internet (ou ailleurs):

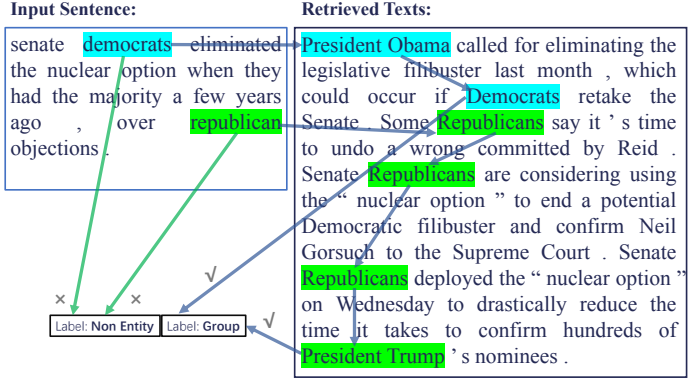
Azawad reprend les armes

Le Mouvement national de l'Azawad (MNA), créé en novembre 2010

Le secrétaire général du mouvement est Ahmed Ould Sidi Mohamed

Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?

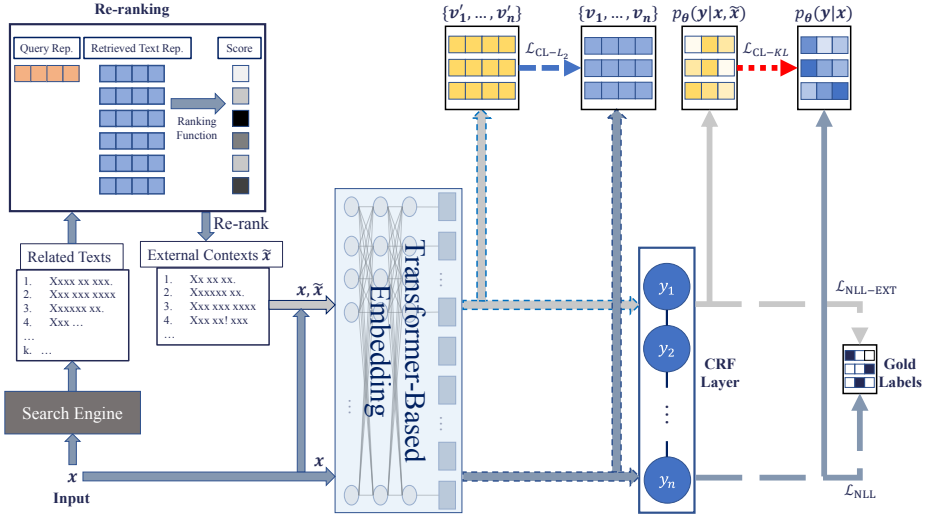


8

[8] Xinyu Wang et al. (2021). "Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning". In: ACL

Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?



Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?

Amélioration des performances: significatives... Mais décevantes

	Social Media		News		Biomedical		E-commerce
	WNUT-16	WNUT-17	CoNLL-03	CoNLL++	BC5CDR	NCBI	
Evaluation: **w/ Context**							
w/ Context	57.43†	60.20†	93.27†	94.56†	90.76†	89.01†	83.15†
CL-L_2	58.61†	60.26†	93.47†	94.62†	**90.99**†	89.22†	83.87†
CL-KL	**58.98**†	**60.45**†	**93.56**†	**94.81**†	90.93†	88.96†	**83.99**†

8

[8] Xinyu Wang et al. (2021). "Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning". In: ACL

Contextualisation et modèles de langue

- Modèle de langue = sélection des documents du contexte (BERT-Score)
- Contextualisation directe possible avec un modèle de langue
- Ouverture: reformulation de phrase
- ... Voire recherche directe des entités

Expériences préliminaires: recherche de prompts

Example of prompt	Persona	Reflection pattern	Answer format
Could you provide more information about the entities in the provided text.			
Act as an expert linguist. Could you provide more information about the entities in the provided text. **Provide outputs that an expert linguist would create.**	✓		
Could you provide more information about the entities in the provided text. Moreover, **Please address any potential ambiguities or limitations in your answer in order to provide a more complete and accurate response.**		✓	
Could you provide more information about the entities in the provided text. **You should enumerate your answers as a list of propositions prefixed by a number.**			✓
You act as an expert linguist, Could you provide more information about the entities in the provided text. **Provide outputs that an expert linguist would create.** Moreover, **Please address any potential ambiguities or limitations in your answer in order to provide a more complete and accurate response.** **Provide outputs that an expert linguist would create**.	✓	✓	✓

⇒ Dépassement des résultats de CL-NER[9]

[9] Herserant et al. 2024. En soumission

Contextualisation et modèles de langue

- Modèle de langue = sélection des documents du contexte (BERT-Score)
- Contextualisation directe possible avec un modèle de langue
- Ouverture: reformulation de phrase
- ... Voire recherche directe des entités

Expériences préliminaires: premiers problèmes

Task	Variation	*Empty*	*Denied*	*Fail*	*Correct*
Reformulation	Classic	214 (6.31%)	374 (11.02%)	441 (12.99%)	2365 (69.68%)
	Persona	215 (6.33%)	257 (7.57%)	262 (7.72%)	2660 (78.37%)
	Reflexion pattern	209 (6.16%)	433 (12.76%)	216 (6.36%)	2536 (74.72%)
	Answer format	-	-	-	-
	All	118 (3.48%)	310 (9.13%)	103 (3.03%)	2863 (84.35%)
Named Entity Recognition	Classic	214 (6.31%)	313 (9.22%)	484 (14.26%)	2383 (70.21%)
	Persona	225 (6.63%)	222 (6.54%)	320 (9.43%)	2627 (77.40%)
	Reflexion pattern	221 (6.51%)	328 (9.66%)	273 (8.04%)	2572 (75.78%)
	Answer format	-	-	-	-
	All	134 (3.95%	258 (7.60%)	109 (3.21%)	2893 (85.24%)
Context Variation	Classic	237 (6.98%)	347 (10.22%)	415 (12.23%)	2395 (70.57%)
	Persona	221 (6.51%)	285 (8.40%)	256 (7.54%)	2632 (77.55%)
	Reflexion pattern	209 (6.16%)	338 (9.96%)	215 (6.33%)	2632 (77.55%)
	Answer format	-	-	-	-
	All	136 (4.01%)	292 (8.60%)	91 (2.68%)	2875 (84.71%)

Conclusion

- Auto-supervision
- Multi-modalité
- Dynamicité + encodage de la position
- Technique d'optimisation
- Contextualisation

- Gangner en performances en NER est difficile

  Et publier en NER est encore plus difficile!
- 100% de performance n'est pas un objectif réaliste