

L'Informatique face à la complexité du langage naturel

Jean Rohmer

14 Juin 2021

Groupe Emergence

Données Structurés , Données non structurées ...

80% des données sont ...

Données Structurés , Données non structurées ...

80% des données sont ... Hyperstructurées !

Ce sont les données en langage naturel

Le langage naturel est ***trop*** structuré pour les ordinateurs

La puissance du langage naturel s'exprime par contraste avec les dizaines de milliers de hiérarchies de menus dans un système Windows Microsoft.

Ce que l'on dit naturellement, la machine peine à le comprendre et le faire simplement

C'est parce qu'il a mis 100000 ans à se construire que le langage naturel se comprends à la vitesse du son

Les données, c'est du langage naturel

« Vers 8h50, le temps au-dessus de la Région Parisienne est faiblement nuageux »

L'ordinateur n'a pas été fait pour traiter les données, mais les nombres

(Exposé Grec-O JR Juin 2015)

« Comme tout bon informaticien, j'ai horreur des nombres »

Alain Colmerauer, France Culture, 1991

Exemples d'application du TALN (Traitement Automatique du Langage Naturel)

- le moteur de recherche Google
- la traduction automatique
- la conversion voix <-> texte

Tout a commencé dès les années 1950

Il se passe depuis quelques années en TALN des choses absolument choquantes:

L'efficacité des très grands modèles de langages.

Avec un chiffre-clé: ***mille milliards***

LA THEORIE DES DEUX TROUS NOIRS

« *dispute* »

LITTERATUS Black Hole
Realm of DOCUMENTS

LAWYERS

WEB 2.0

KM

| **ZONE INTERDITE** |

Object Oriented Languages

Semantic WEB

SQL

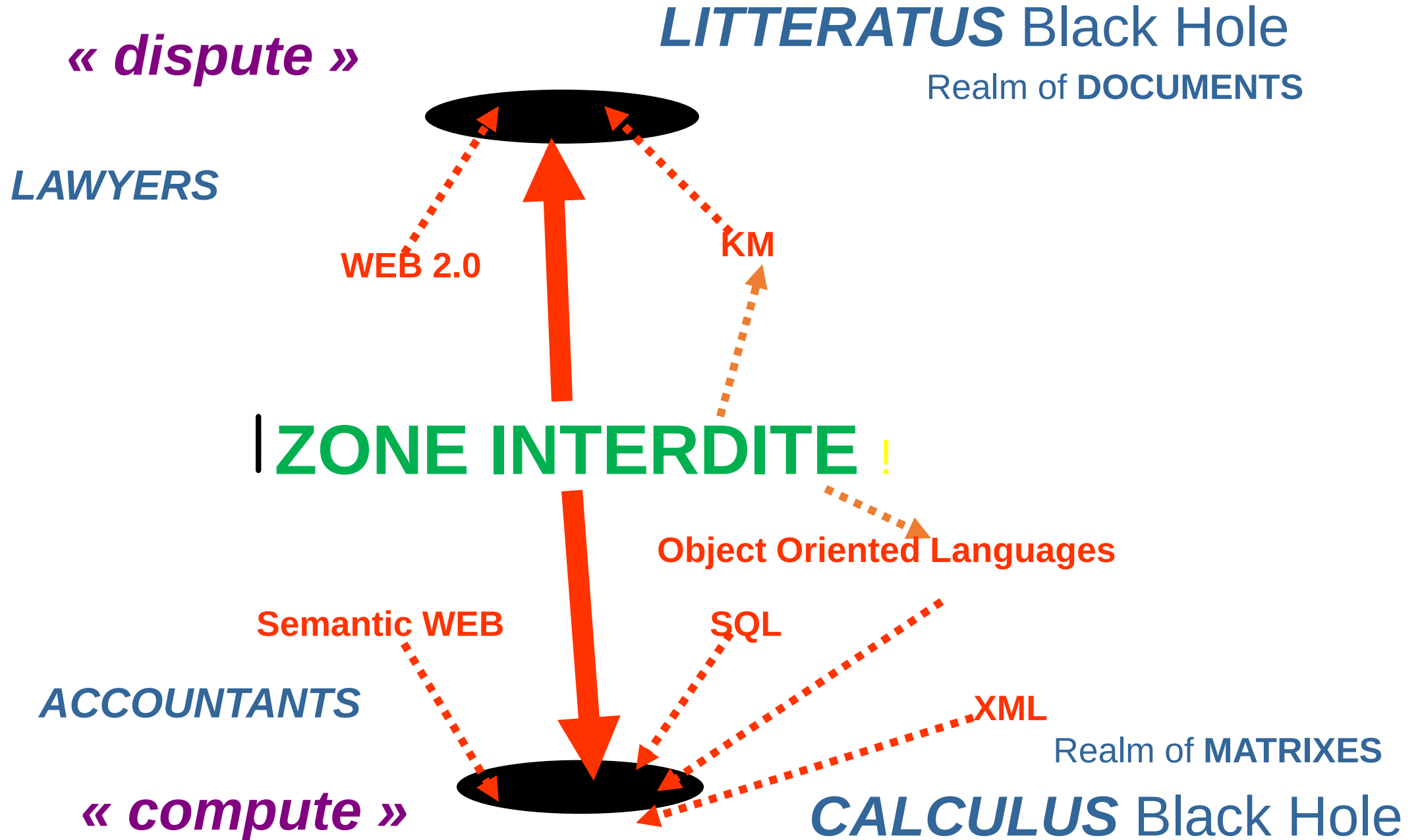
ACCOUNTANTS

XML

Realm of MATRIXES

« *compute* »

CALCULUS Black Hole



DISPUTE

LITTERATURE

PHILOSOPHIE

LINGUISTIQUE

MODELISATION

PROGRAMMATION

MATHEMATIQUES

CIRCUITS ELECTRONIQUES

PHYSIQUE

COMPUTE

COMMENT OCCUPER LA ZONE INTERDITE ?

IDEE D'UN LANGAGE INTERMEDIAIRE

- « établir un ordre entre toutes les pensées, ... de même qu'il y en a un établi entre les nombres »
- « cette langue aiderait au jugement , lui représentant si distinctement les choses qu 'il lui serait presque impossible de se tromper »
- « je tiens que cette langue est possible ... mais n 'espérez jamais la voir en usage ... sauf au Paradis Terrestre ... »

Descartes

- « établir un ordre entre toutes les pensées, ... de même qu'il y en a un établi entre les nombres »
- « cette langue aiderait au jugement , lui représentant si distinctement les choses qu 'il lui serait presque impossible de se tromper »
- « je tiens que cette langue est possible ... mais n 'espérez jamais la voir en usage ... sauf au Paradis Terrestre ... »

- « quoique cette langue dépende de la vraie philosophie, elle ne dépend pas de sa perfection »
- « à mesure que la science des hommes croîtra, cette langue croîtra aussi »
- « alors raisonner et calculer sera la même chose »

Leibniz

- « quoique cette langue dépende de la vraie philosophie, elle ne dépend pas de sa perfection »
- « à mesure que la science des hommes croîtra, cette langue croîtra aussi »
- « alors raisonner et calculer sera la même chose »

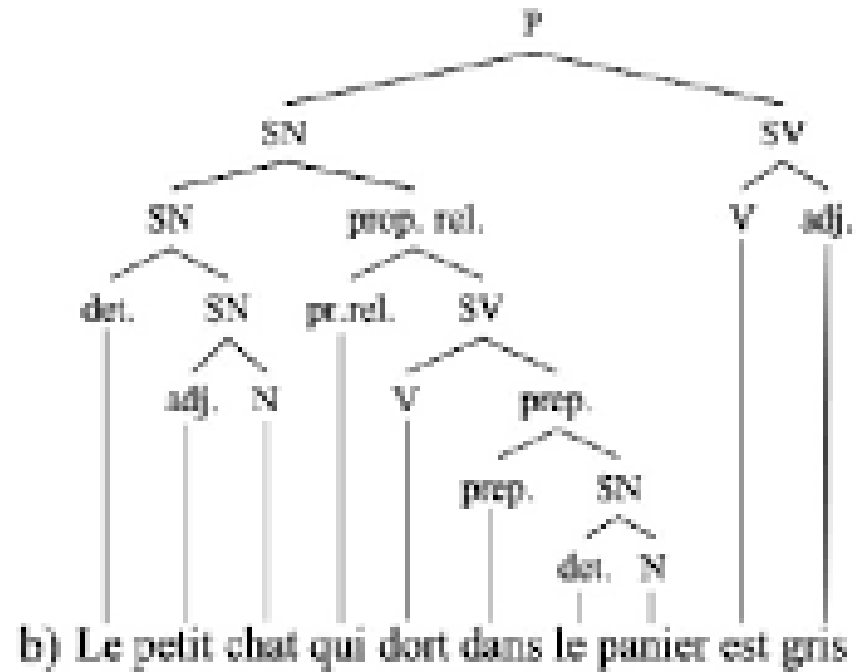
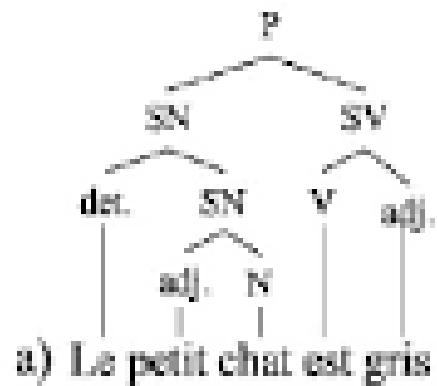
L'INTELLIGENCE ARTIFICIELLE SYMBOLIQUE ESSAYE
BRAVEMENT D'OCCUPER LA ZONE INTERDITE

L'INTELLIGENCE ARTIFICIELLE MODERNE VEUT
PRECIPITER VIOLEMMENT LE LANGAGE DANS LA
MACHINE

COMMENT OCCUPER LA ZONE INTERDITE ?

IDEE D'UN LANGAGE INTERMEDIAIRE

EXEMPLE: L'ANALYSE GRAMMATICALE



Approche Symbolique: Prolog DCG

```
sentence(s(NP,VP)) --> noun_phrase(NP), verb_phrase(VP).
noun_phrase(np(D,N)) --> det(D), noun(N).
verb_phrase(vp(V,NP)) --> verb(V), noun_phrase(NP).
det(d(the)) --> [the].
det(d(a)) --> [a]. noun(n(bat)) --> [bat].
noun(n(cat)) --> [cat].
verb(v(eats)) --> [eats].
```

One can now query the interpreter to yield a parse tree of any given sentence:

```
| ?- sentence(Parse_tree, [the,bat,eats,a,cat], []).
Parse_tree = s(np(d(the),n(bat)),vp(v(eats),np(d(a),n(cat)))) ? ;
```

Une langue: Milliers de règles assez complexes, dictionnaires complets

Grosses impasses: ponctuation, indentation, séquences, parenthèses ...

Résultats pas terribles ...

EUROTRA Début des années 1980:

Aerospatiale construit des satellites européens

Aerospatiale builds european satellites

Résultats pas terribles ...

EUROTRA Début des années 1980:

Aérospatiale construit des satellites européens

Aérospatiale builds european satellites

Aérospatiale construit les satellites européens

Aérospatiale builds european satellites

NB: Traduction Espagnol Danois

UN GRAND MYSTERE POUR LES INFORMATICIENS:

LES LINGUISTES !

"Every time I fire a linguist, the performance of the
speech recognizer goes up"

Fred Jelinek, IBM, 1998

UN GRAND MYSTERE POUR LES INFORMATIENS: LES LINGUISTES !

« Un linguiste c'est un parti, deux linguistes c'est une
tendance, trois linguistes c'est une scission. » Lénine.

Exemple d'infini de la linguistique : livre de Bertil
Malmberg

La clé de ma réussite ?

Linguistics ! NO Linguistics! Winston Churchill

Les informaticiens ne comprennent pas pourquoi les linguistes utilisent aussi peu de formalismes pour décrire leurs théories.

Il utilisent essentiellement ... le langage naturel, en inventant sans arrêt des termes et des concepts, et en s'étripant constamment entre chapelles.

Liens avec l'art, la littérature, la psychanalyse, la philosophie, les sciences cognitives, histoire, géographie, biologie, physiologie, acoustique, vision, neurobiologie, etc, etc,

François Rastier vous expliquera ça.

De Saussure, Jakobson, Peirce, Harris, Chomsky, Tesnière

Notion de différentiation, opposition

Peirce: le sens d'un symbole est sa traduction en d'autres symboles

La double articulation: système de combinaison des sons, système de
combinaison des sens

Un « petit » problème « simple »:

Dans un paragraphe, délimiter les phrases.

« Je suis allé à Paris. Il faisait beau. »

Approche symbolique: des règles:

Règle 1: Un « . » suivi d'un blanc, suivi d'une majuscule.

Mais « Mr. Louis Martin habite à Paris »

Mais « Paul hésitait entre Dr et Mr. Louis n'hésitait pas. »

Mais « Il est venu avec ... Marie »

Mais « Il a retrouvé son parapluie, donc il est venu avec ... Il est arrivé tard.

Approche par apprentissage: on part d'exemples.

Apprentissage supervisé: on donne explicitement des exemples

« Je suis allé à Paris. // Il faisait beau. »

Mais « Mr. Louis Martin habite à Paris »

Mais « Il hésitait entre Dr et Mr. // Louis n'hésitait pas. »

Mais « Il est venu avec ... Marie »

Mais « Il a retrouvé son parapluie, donc il est venu avec ... // Il est arrivé tard.

Approche par apprentissage supervisé: étiquetage humain: lent, couteux, pas fiable

Apprentissage NON supervisé: on se débrouille !

Exemples de fins de phrases: des phrases en fin de paragraphes

Exemples de débuts de phrases: des phrases en début de paragraphes

Pas besoin d'énoncer des règles, pas besoin d'annoter des exemples

On laisse la machine se débrouiller

On peut travailler sur de très grandes masses de données

Ca marche très bien

Mais ... on ne sait pas comment ça marche

On ne sait pas à l'avance ce que ça va donner

On ne sait pas expliquer, justifier, valider, certifier ...

Le CHOC ... des MOTS

Depuis quelques années

Avec les réseaux de neurones profonds

On obtient des résultats

Absolument bluffants avec

Des corpus d'exemples gigantesques

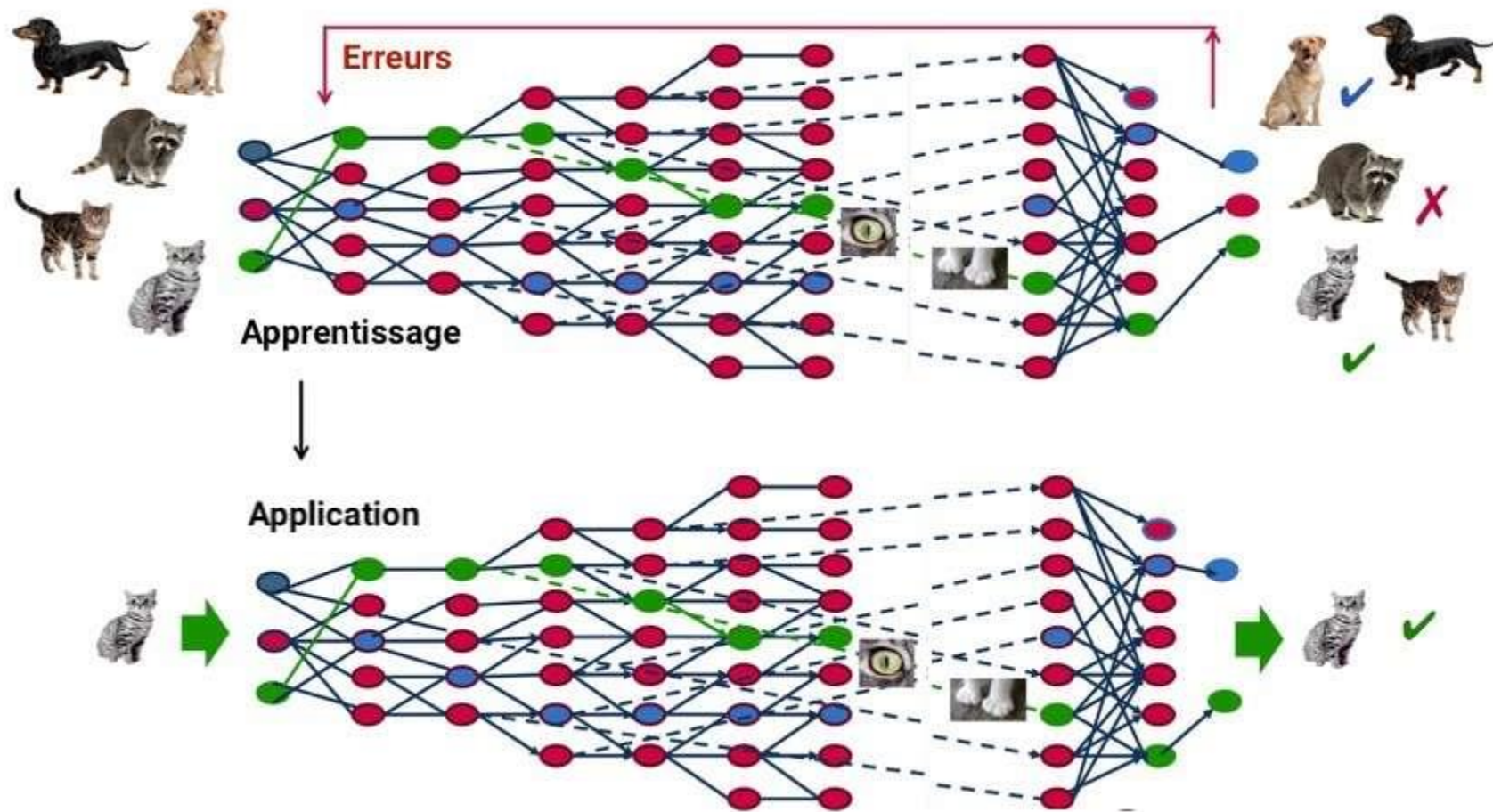
Des machines surpuissantes

Des temps de calcul énormes

Des consommations d'énergie démentes

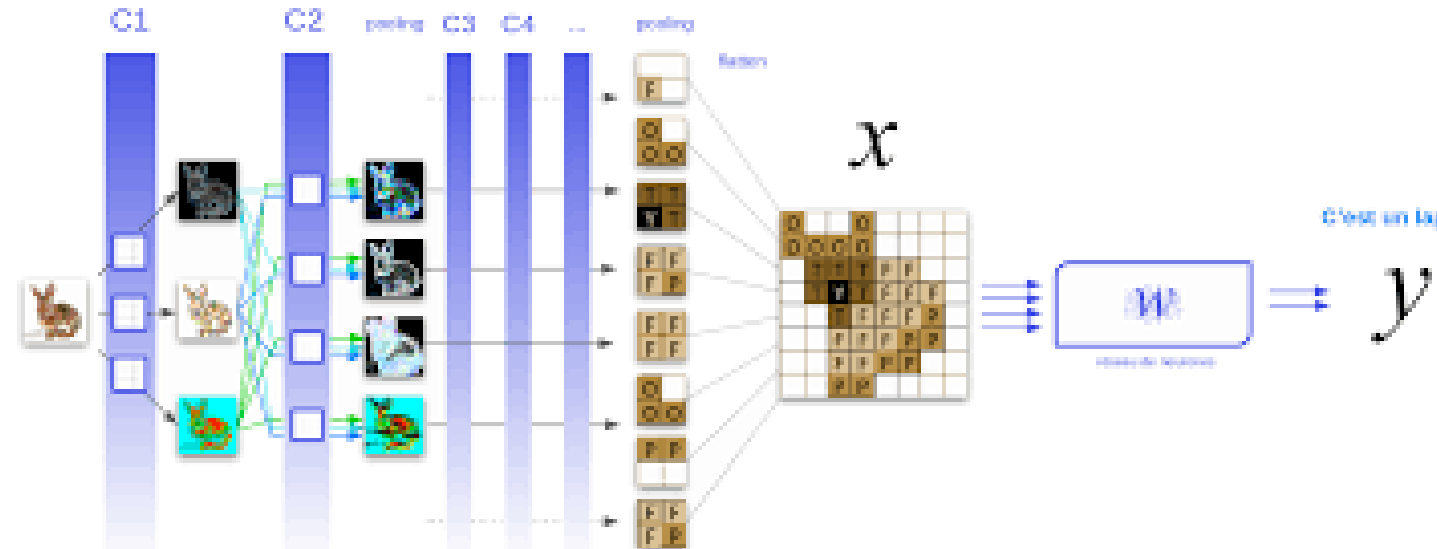
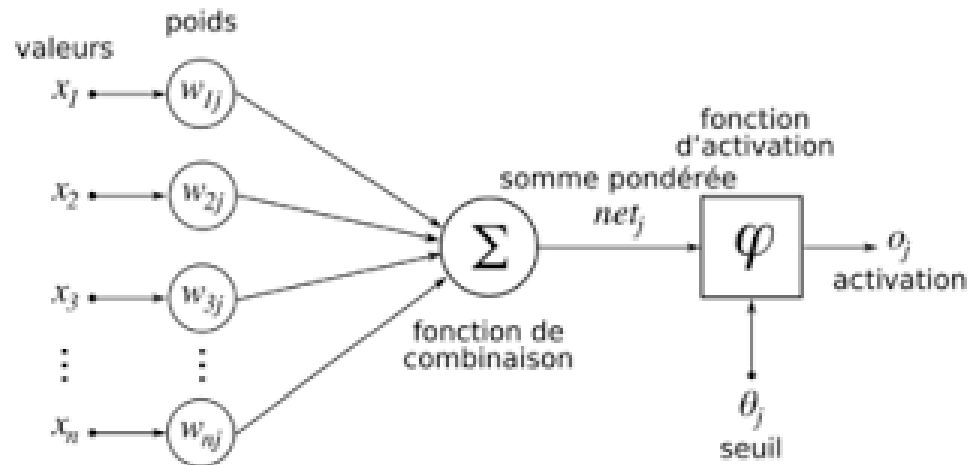
Les rares grosses équipes d'ingénieurs d'élite des GAFAM

... et on sait de moins en moins expliquer comment ça marche



$$\frac{\partial E}{\partial w_{ab}^{(l)}} = \sum_k \frac{\partial x_k^{(n-1)}}{\partial w_{ab}^{(l)}} \sum_i w_{ik}^{(n)} \underbrace{g'^{(n)}(h_i^{(n)})(y_i - t_i)}_{e_i^{(n)}}$$

L'origine: transposer aux textes les prouesses de l'apprentissage profond qui a révolutionné la reconnaissance d'images



Problèmes:

Images:

Ca marche sur des nombres réels

Avec des fonctions continues dérivables

**Pas grave: on va transformer les mots en vecteurs de
nombres!**

A un mot, on va associer un vecteur de centaines ou milliers de nombres

On prend un très gros corpus, on cherche un / des critères de ressemblance entre mots (exemple: ils ont souvent les mêmes voisins)

Et on cherche par optimisation combinatoire des valeurs de chaque vecteur tels que les mots qui se ressemblent aient des vecteurs proches.

En anglais, on dit « embedding », word2vec ...

Chaque mot devient un « pixel de texte » avec 512

« Couleurs Sémantiques »

Et on enfourne dans notre réseau de neurones !

Plus besoin de règles de morphologie, de syntaxe

Les accords genre, nombre, on oublie

Plus besoin de dictionnaires

Plus besoin de linguistes

Le langage naturel: un rêve pour l'apprentissage non supervisé:

Apprendre la fonction, qui à toute suite de N mots, donne la probabilité des mots suivants:

« Le Modèle du Langage »

Une guerre permanente entre GAFAM et Chinois depuis bientôt 10 ans

Le petit dernier!

<https://arxiv.org/abs/2104.12369>

Des réseaux de neurones à 1000 milliards de paramètres

Des corpus d'apprentissage de 1000 milliards de caractères

Des machines approchant l'Exaflop par seconde

1 milliard de milliard d'opérations par seconde

Des durées d'apprentissage de 10 jours ou plus

Une colossale dépense d'énergie

Enormément de cuisine de réglage, dont les détails sont rarement publiés

INCROYABLES EXPLOITS EN PRODUCTION AUTOMATIQUE DE TEXTE

<https://arr.am/2020/07/09/gpt-3-an-ai-thats-eerily-good-at-writing-almost-anything/>

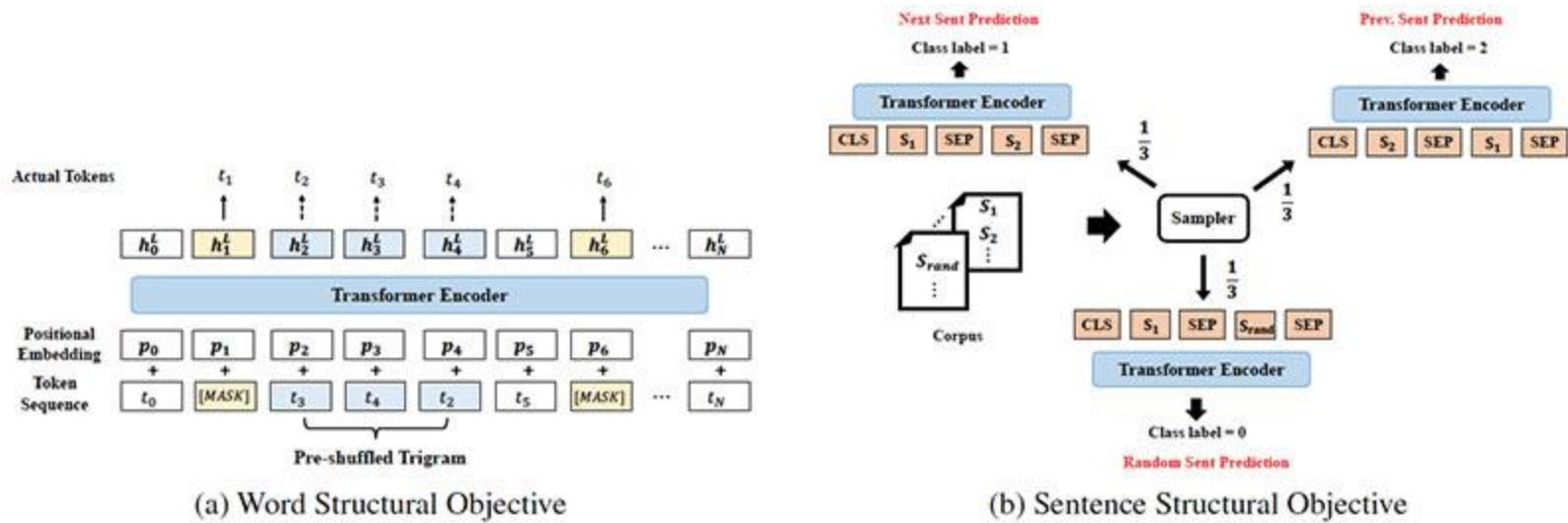


Figure 1: Illustrations of the two new pre-training objectives

Environ tous les 3 mois, une annonce qui « Outperforms Competition »

Training a typical natural-language processor requires so much computing power that it emits as much carbon as the life span of five American cars. Training an image recognition model releases as much energy as a typical home puts out in two weeks—and it's something that leading tech companies do multiple times a day

Principe de Base (En traduction par exemple)

Au départ, A, B et C sont des fonctions de calcul avec des paramètres
(« 1000 milliards » en tout) initialisé au hasard

On présente des milliards de phrases A et Phrases B par couple

On ajuste à chaque fois un peu la valeur des paramètres



Et on repasse sur GOOGLE TRANSLATE en 2021 notre exemple qui ne marchait pas en 1985:

Et on repasse sur GOOGLE TRANSLATE en 2021 notre exemple qui ne marchait pas en 1985:

AIRBUS construit des satellites européens

AIRBUS builds european satellites

AIRBUS construit les satellites européens

AIRBUS builds european satellites

On arrive même à faire de la traduction SANS couples de phrases alignée.

On donne juste des ensembles de textes dans des langues différentes qui parlent plus ou moins du même sujet

A propos d'une question de G. Balantzian

Les systèmes d'information, c'est aussi complexe que le langage naturel, mais là, on n'a pas des milliards d'exemples ...

CONCLUSION:

**Travailler un an en Intelligence Artificielle vous convainc de l'existence
de Dieu
(Alan Perlis)**

CONCLUSION:

**Travailler un an en Intelligence Artificielle suffit à vous convaincre de
l'existence de Dieu
(Alan Perlis)**

**Travailler un mois en Linguistique suffit à vous convaincre de
l'existence du Diable
(JR)**