Notre sujet

TAL

Apprentissage pour le TAL

Réseaux de neurones

Apprentissage profond

Modèles de langues

Giga-modèles

Giga Modèles ?

Anaïs Halftermeyer

LIFO - Université d'Orléans

RITUEL Mars 2023

Notre sujet

TAL

Apprentissage pour le TAL

Réseaux de neurones

Apprentissage profond

Modèles de langues

Giga-modèles

Notre Sujet

Le groupe de travail RITUEL
"Recherche d'Information et Traitements **Utiles et Eclairés** des Langues"
https://www.info.univ-tours.fr/RITUEL/
dans le cadre du réseau thématique de recherche

RTR DIAMS
RÉSEAU THÉMATIQUE DE RECHERCHE
Données, Intelligence Artificielle, Modélisation et Simulation

Notre sujet

TAL

Apprentissage pour le TAL

Réseaux de neurones

Apprentissage profond

Modèles de langues

Giga-modèles

Notre Sujet

Présentation (très) générale des Large Language Models ou giga-modèles

Quel objectif ?

Produire des outils pour traiter la langue :

- traduire
- résumer
- extraire de l'information
- répondre à des questions
- résoudre des coréférences
- analyser syntaxe
- comprendre ?
- ...

Méthodes

A l'origine des méthodes symboliques : on rédige des grammaires, on cherches des représentations formelles pour édicter les règles sous-jacentes de fonctionnement du "système" langue lorsque mis en oeuvre dans le discours : le **TAL théorique**.

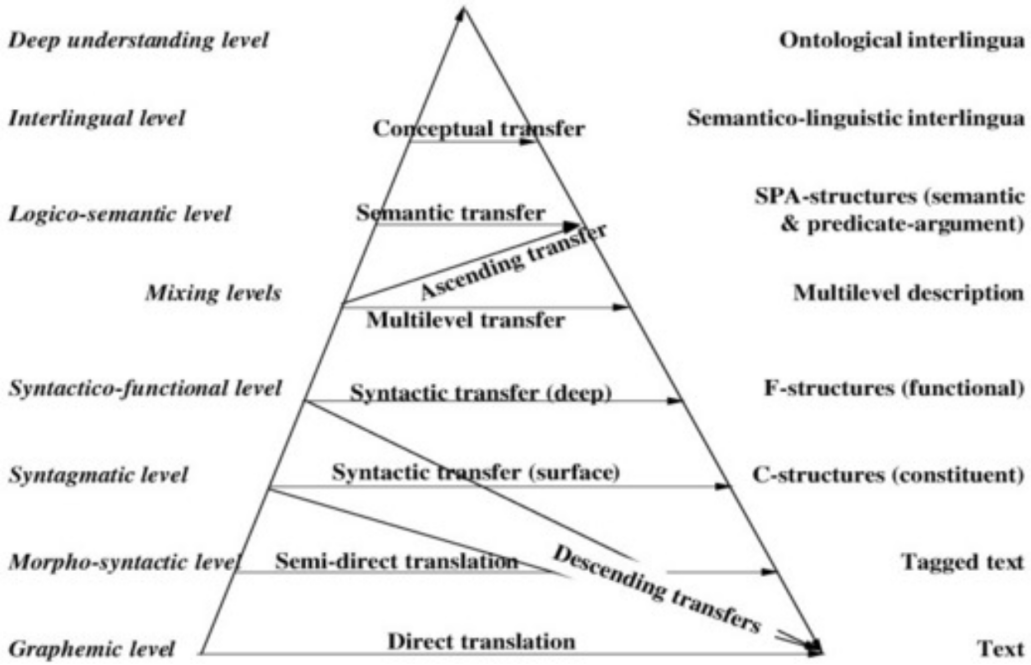
Puis les révolutions technologiques successives ont permis de proposer des méthodes numériques, dite aussi méthodes stochastiques, pour un **TAL robuste**.

TAL théorique ou TAL robuste peuvent s'attaquer uniquement à un niveau d'analyse ou être au coeur de traitements plus complets. (Cori Marcel, Des méthodes de traitement automatique aux linguistiques fondées sur des corpus dans Languages 2008)

Architecture

Voici un exemple d'architecture complète pour de la traduction automatique selon Boitet Christian "Automated Translation" dans Revue française de linguistique appliquée 2003 :

Figure 1 – Vauquois' triangle



Apprentissage

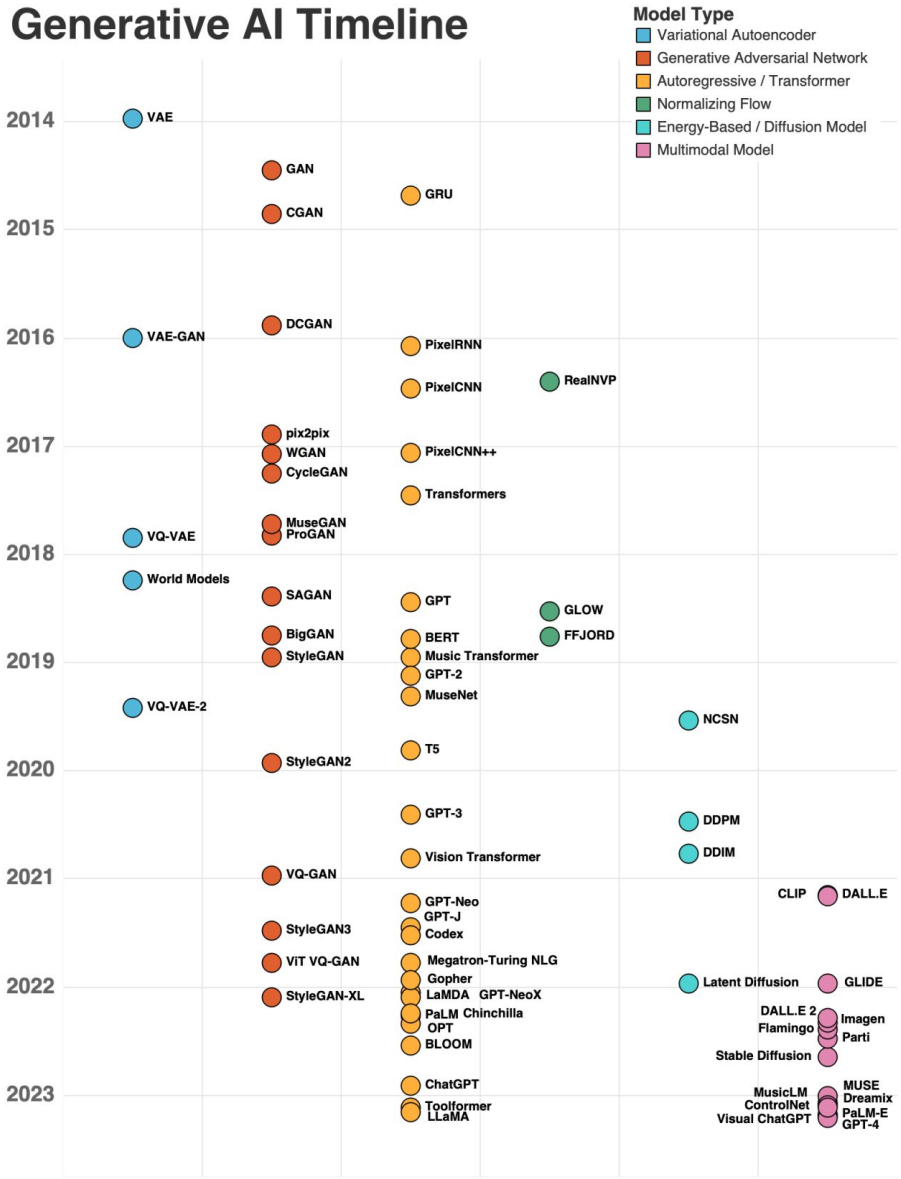
Les approches numériques ont donné naissance à une grande variété de méthodes d'apprentissage automatique, parmi elles :

- apprentissage supervisé = on montre les réponses attendues pour un certain nombre d'exemples
- apprentissage non-supervisé = on cherche à regrouper les données sans connaître a priori les groupes attendus

les approches biomimétiques

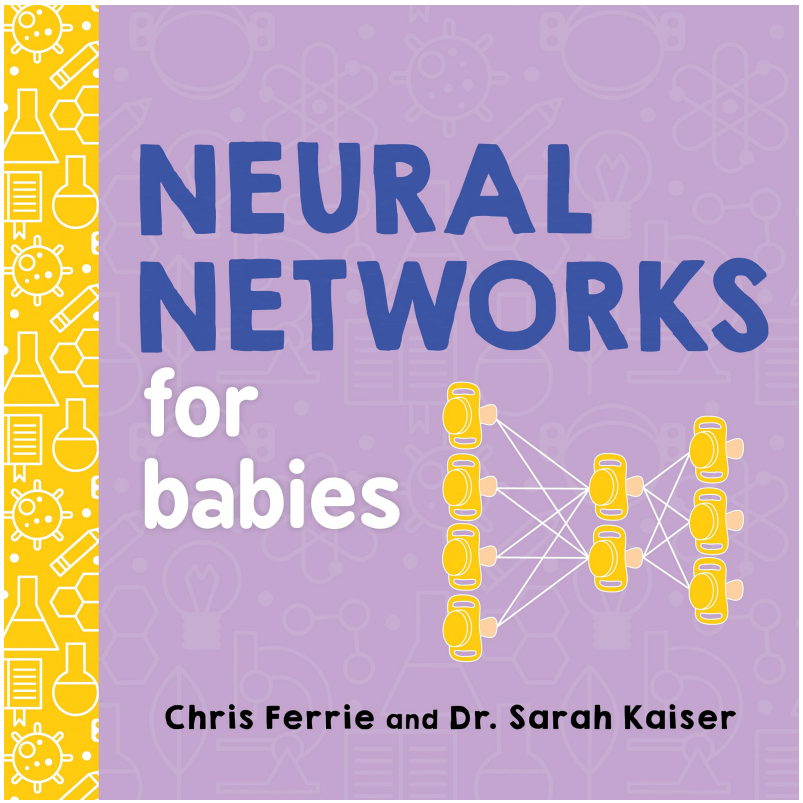
Les propositions de réseaux de neurones sont assez anciennes mais ont été un peu malmenées par le passé faute de puissance de calcul et de mémoire pour les mettre à profit jusqu'à... il y a peu !

- 1881 : théorisation des neurones humains par Heinrich Wilhelm Waldeyer
- 1943 : modèle du neurone formel par Warren McCulloch et Walter Pitts
- 1957 : premier perceptron (une couche d'entrée, une couche de sortie) produit par Frank Rosenblatt
- 1969 : arrêt brutal des avancées sur ce sujet suite à *Perceptrons* de Marvin Lee Minsky et Seymour Papert
- 1982 : nouveau modèle de John Joseph Hopfield
- puis...



tweet de David Foster 03/01/2023

Réseaux ? neurones ?



Notre sujet

TAL

Apprentissage pour le TAL

Réseaux de neurones

Apprentissage profond

Modèles de langues

Giga-modèles

Lien entre LLM et tranformers

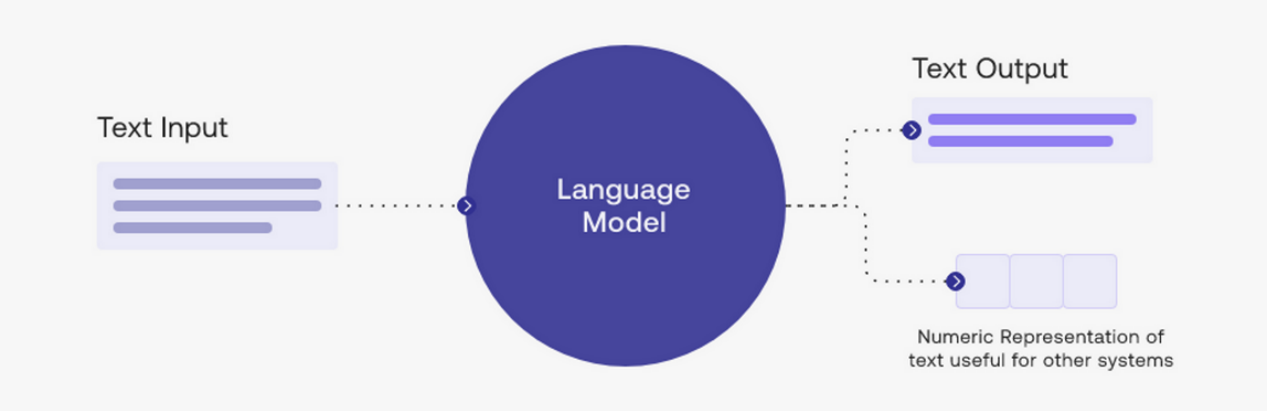
Les architectures tels les réseaux de neurones récurrents (RNN, dont les LSTM par exemple ou encore les GRU) ont répandu leur usage, mais traitent le texte de manière séquentielle.

Les architectures **transformers** ont fait leur apparition permettant de paralléliser et de prendre en entrée le texte d'un seul coup.

Modèles pré-entraînés

Les modèles de langues sont des modèles qui ont la capacité de prédire la probabilité d'**une séquence de mots** sans supervision.

Ils ont la spécificité d'avoir été **pré-entraînés** sur une certaine quantité de données.

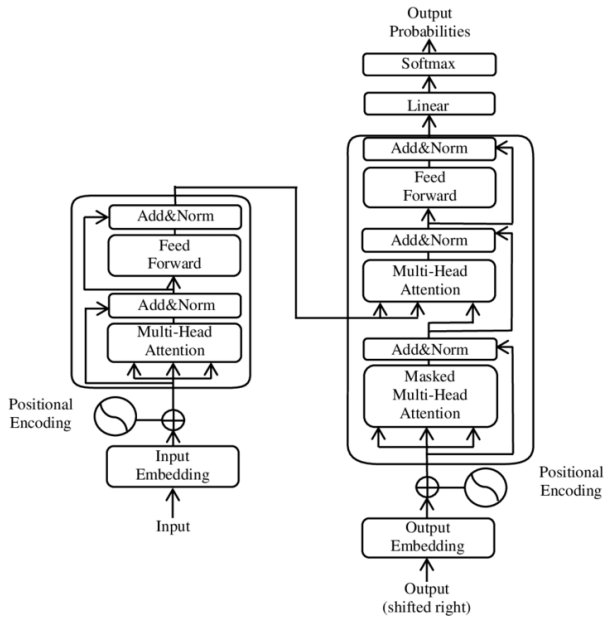


(tiré du site The Cohere Platform)

Ils acquièrent une **représentation de la langue**.
Ils peuvent ensuite être "*fine-tunés*" de manière supervisée sur des taches plus spécifiques.

On peut se représenter les giga-modèles comme des couches multiples de réseaux de neurones :

- Couches d'*embedding* (plongement) : convertit la séquence de mots en entrée en une représentation vectorielle de grande dimension
- *Encoder-decoder* : incorpore des informations contextuelles sur chaque token
- Couches *feedforward* : plusieurs couches complètement connectées qui appliquent des transformations non linéaires aux plongements
- Méchanisme d'attention : permet de se focaliser sur certaines parties des données pour optimiser le traitement



Exemple d'architecture tirée de Wikipédia

Notre sujet

TAL

Apprentissage pour le TAL

Réseaux de neurones

Apprentissage profond

Modèles de langues

Giga-modèles

Giga ?

Basés sur les transformers, leur architecture leur permet de disposer de "**self-attention**".

Ces gains techniques ont permis l'accession au statut de "large" ou giga en étendant à des **millions de paramètres autrement dit de neurones** pour l'apprentissage de ces modèles.

La "**profondeur**" atteinte par ces architectures et les avancées techniques notamment concernant l'attention ont permis des représentations de plus en plus **abstraites** des données étudiées.

La question se pose donc de **la représentation interne de la langue** obtenue par ces outils dont on ne peut exhiber la représentation cachée !