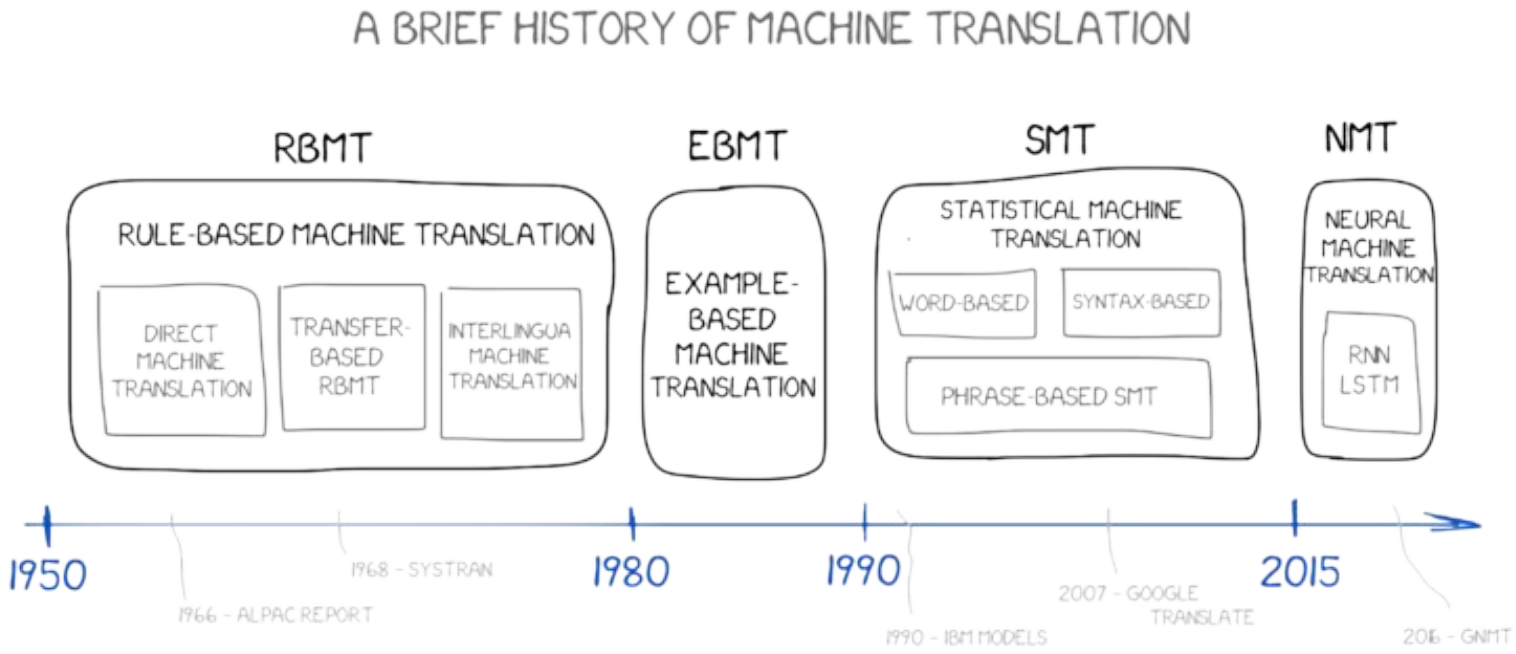
Approches neuronales pour la traduction automatique, historique, situation et enjeux

Colloque éthique et traduction, Février 2023

Fabrice Lefèvre; Laboratoire d'Informatique d'Avignon, Avignon Univ.

Histoire de la SMT

- Stochastic machine learning
  - vs symbolic (Rule-Based MT)
- Recours aux modèles probabilistes, basé sur de **larges corpus de bi-textes**
  - Déjà ancien
  - SotA Phrase-based MT



source: freeCodeCamp

Pros and cons

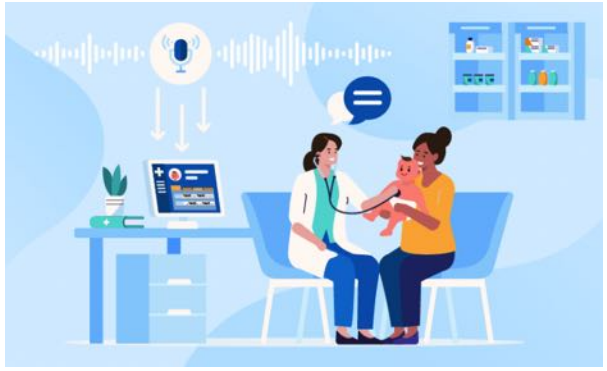
- Avantages
  - Gestion des volumes, productivité
  - Disponibilité, rapidité
  - Réduction des coûts en usage
  - Constance

- Limites
  - Erreurs (naïves), imprécisions
  - Prise en compte du sens faible
  - Coût élevé de la post-édition manuelle
  - Coût élevé de l'adaptation au domaine, à une langue

AVIGNON UNIVERSITÉ

Difficultés variables

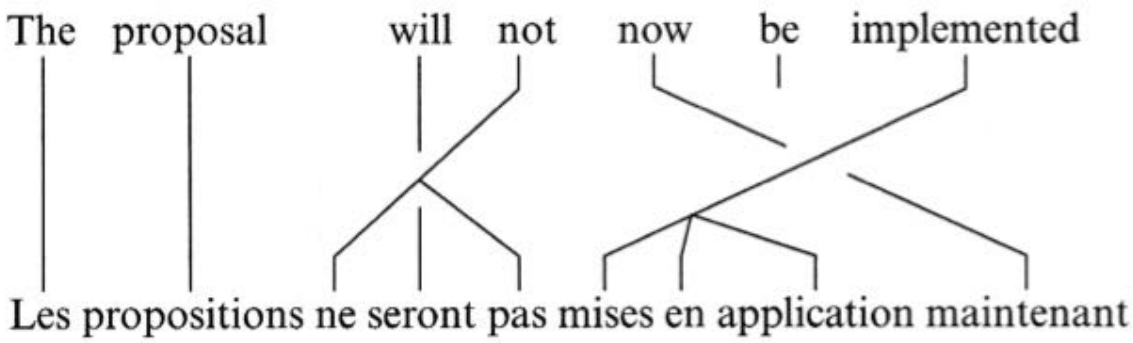
- Différentes tâches/protocoles/cadres applicatifs, différents équilibres :
  - Documents
    - ++ : temps long, post-édition
    - -- : longueur du texte, langage complexe, sémantique élaborée
  - Spontanée (live/on-the-fly)
    - ++ : texte court, langage plus commun, plus prévisible
    - -- : immédiateté de la réponse, pas (ou peu) de corrections
  - Parole (vs texte)
    - -- : latence transcriptions, erreurs de transcriptions, phénomènes de l'oral (hésitations, reprises, agrammaticalité…)



AVIGNON UNIVERSITÉ

Quelques problèmes à résoudre

- Lexique parallèle (polysémie, ambiguïtés…)
- Alignement
  - Calculabilité (complexité NP)
  - Prise en compte des syntagmes (*Phrase-based*)
  - Forte variabilité entre langues, selon typologie syntaxique notamment
  - Correspondance lexicale (contextualisée)
    - Fertilité (0 ou multi)
- Les fameux IBM models (1 à 4) !



AVIGNON UNIVERSITÉ

Quelques problèmes à résoudre

- Lexique parallèle (polysémie, ambiguïtés…)
- Alignement
  - Did you see the look on her **face**$_1$?
  - We could see the clock **face**$_2$ from below
  - It could be time to **face**$_3$ his demons
  - There are a few new **faces**$_4$ in the office today

Face 1

- The most important side (of the head)
- Represents you / yourself
- Used to inform / communicate
- Points forward when you address/confront something

Face 2

- The most important side (of the head)
- Used to inform / communicate

Face 4

- Represents you / yourself

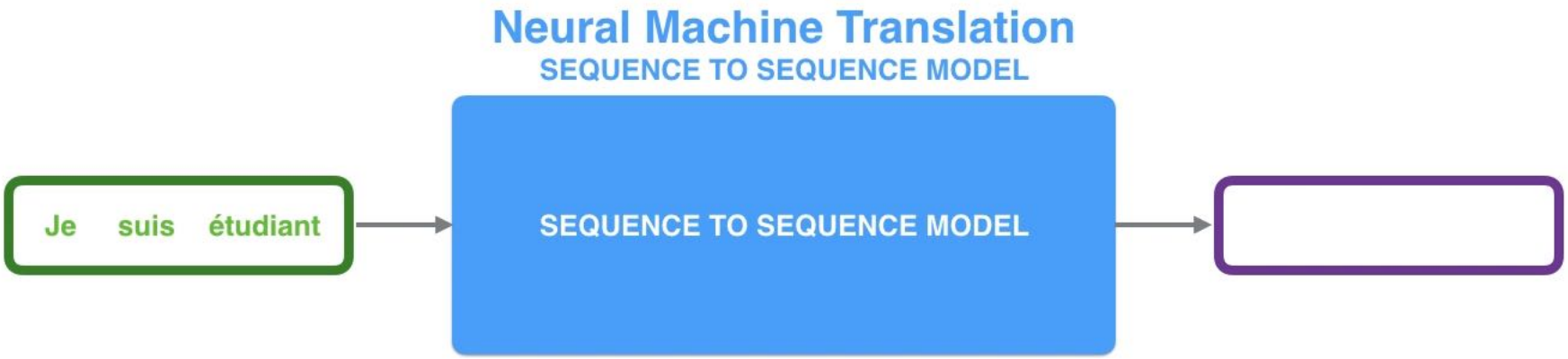
Face 3

- Points forward when you address/confront something

- Le

Les propositions ne seront pas mises en application maintenant

Approche plus holistique

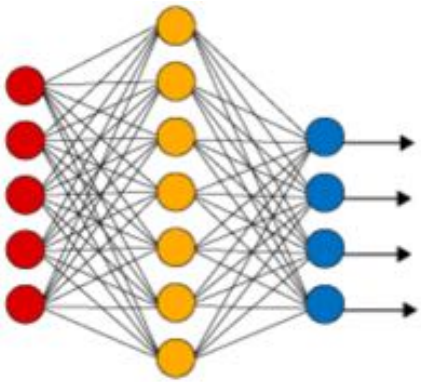


source: http://jalammar.github.io/illustrated-transformer/

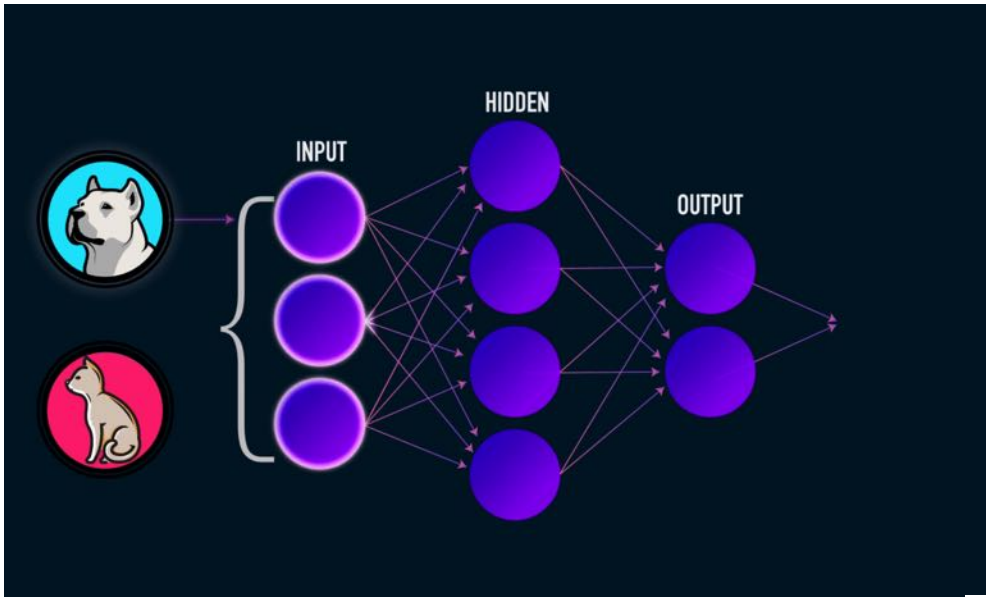
Neural Machine Translation…

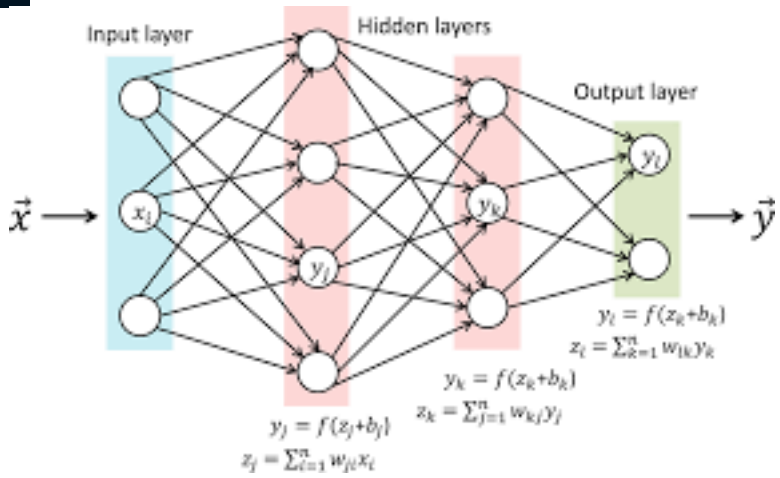
- Approche très ancienne (préhistoire de l'IA, ~1950)

- Dès l'engouement pour le Deep Learning (~2010)
  - Deep Learning = version qui marche des réseaux de neurones (synonymes ci-après) qui empile beaucoup de couches de neurones (deep)



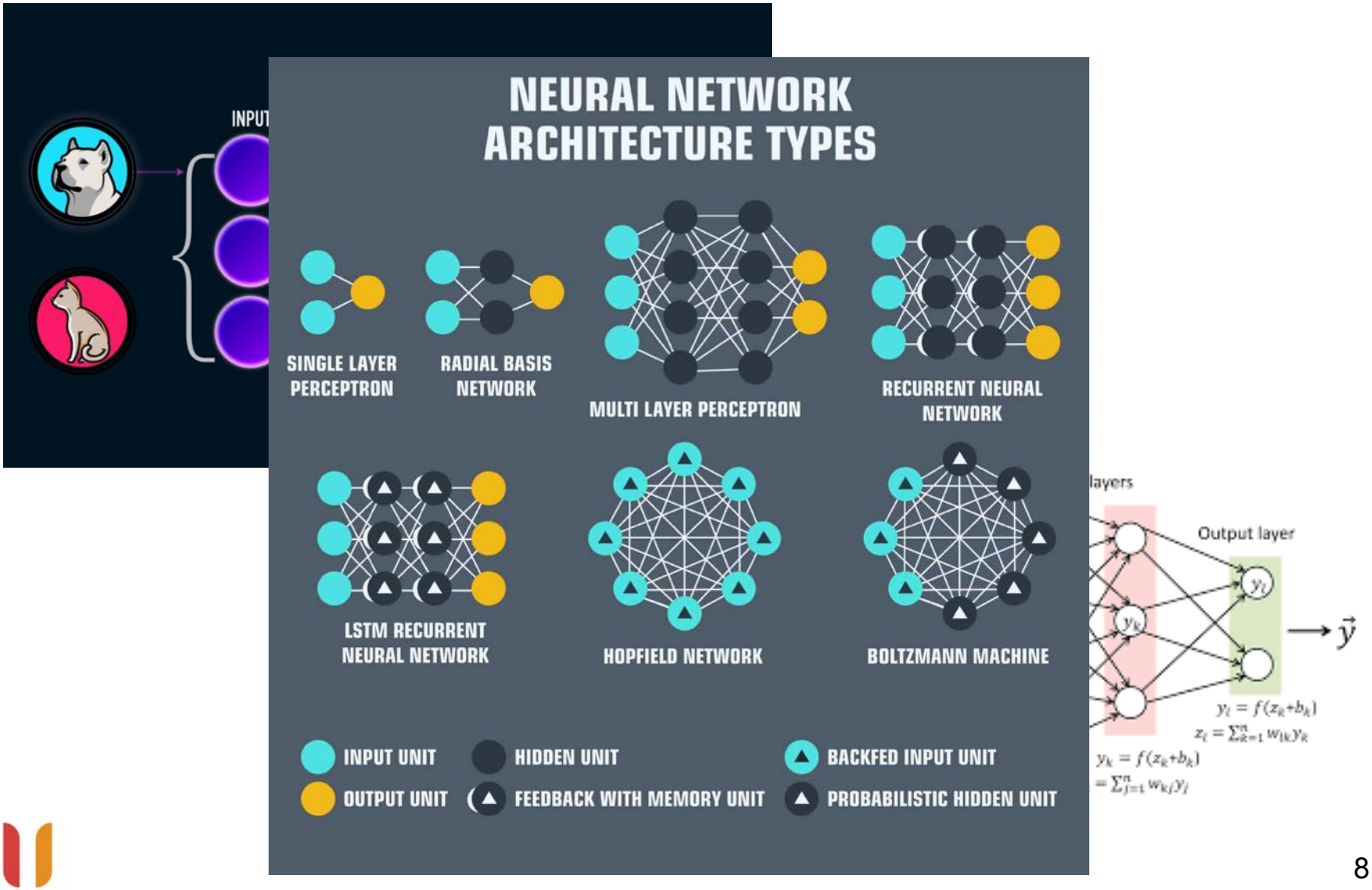
Réseau de neurones artificiels





AVIGNON UNIVERSITÉ

Réseau de neurones artificiels



Neural Machine Translation…

- Approche très ancienne (pré-histoire de l'IA, ~1950)

- Dès l'engouement pour le Deep Learning (~2010)
  - Deep Learning = version qui marche des réseaux de neurones (synonymes ci-après) qui empile beaucoup de couches de neurones (deep) ➔ rappel rapide NN (estimateur de fonctions !)

  - Atout majeur :
    - espace de représentation continu (vectoriel)

AVIGNON UNIVERSITÉ

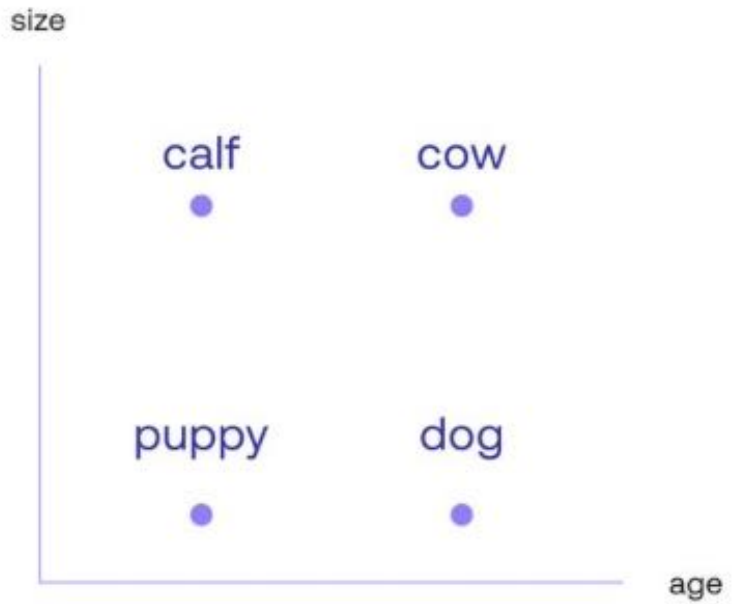
Word embeddings

- Principe des espaces de représentation continus
  - Part d'un vecteur binaire 1-hot : taille du lexique avec un 1 sur l'index du mot correspondant
    - En pratique on utilise des tokens (pas les mots directement) – Byte-Pair Encoding par exemple
  - Projeté vers un espace dont les dimensions ont été apprises AUTOMATIQUEMENT pour représenter des caractéristiques possibles, communes entre les mots
    - **Analyse distributionnelle (co-occurrences,…)**
    - Propriétés topologiques associées à des propriétés syntaxiques ou sémantiques
      - E('Queen') – E('Woman') = E('King') – E('Man')
      - E('France') – E('Paris') = E('Spain') – E('Madrid')

    → TOUT L'ENJEU DE LA PHASE D'ENTRAINEMENT
  - Plongement des mots dans cet espace (« word embedding »)

Word embeddings

- Principe des espaces de représentation continus
  - Part d'un vecteur binaire 1-hot : taille du lexique avec un 1 sur l'index du mot correspondant
    - En prat … rectement) – Byte-Pair Encodi…
  - Projeté vers … oprises AUTOMATI… tiques possibles, communes
    - **Analys…**
    - Proprié… és syntaxiques ou sémant…
      - E('… n')
      - E('… d')



→ TOUT L'ENJEU DE LA PHASE D'ENTRAINEMENT

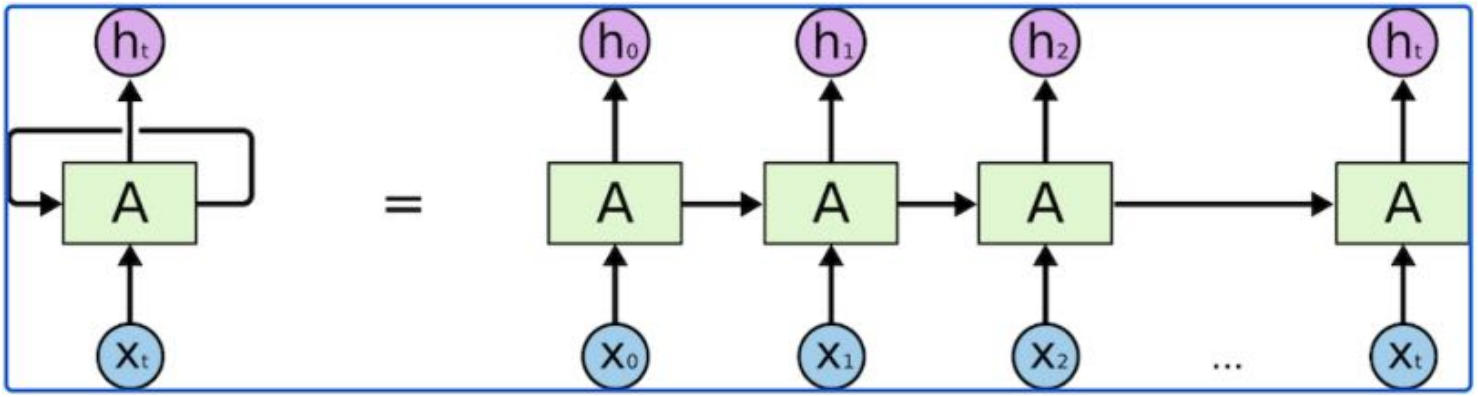
  - Plongement des mots dans cet espace (« word embedding »)

Neural Machine Translation

- Approche très ancienne (pré-histoire de l'IA, ~1950)
- Dès l'engouement pour le Deep Learning (~2010)
  - Deep Learning = version qui marche des réseaux de neurones (synonymes ci-après) qui empile beaucoup de couches de neurones (deep) → rappel rapide NN (estimateur de fonctions !)
  - Atout majeur : espace de représentation continu (vectoriel)
  - Mais gestion de séquences → récurrence,
    - Rétro-propagation du gradient : très lourd à calculer (problème du vanishing gradient)

AVIGNON UNIVERSITÉ

Récurrence des neurones



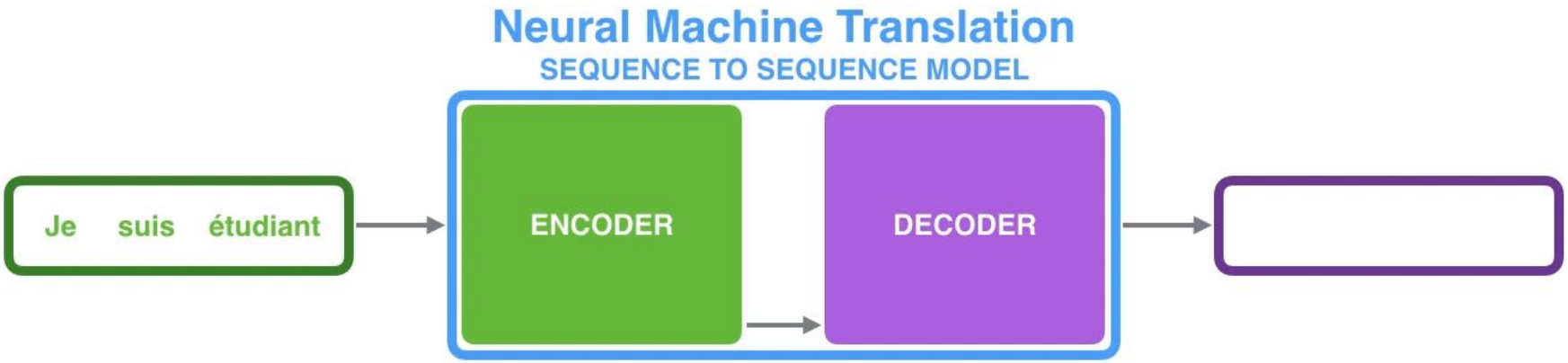
source: http://jalammar.github.io/illustrated-transformer/

Neural Machine Translation

- Approche très ancienne (pré-histoire de l'IA, ~1950)
- Dès l'engouement pour le Deep Learning (~2010)
  - Deep Learning = version qui marche des réseaux de neurones (synonymes ci-après) qui empile beaucoup de couches de neurones (deep) ➔ rappel rapide NN (estimateur de fonctions !)
  - Atout majeur : espace de représentation continu (vectoriel)
  - Mais gestion de séquences → récurrence,
    - Rétro-propagation du gradient : très lourd à calculer (problème de vanishing gradient)
- Modèles Séquence-to-Séquence
  - La bonne idée !
  - Découplée encodage de l'entrée et décodage de la sortie

AVIGNON UNIVERSITÉ

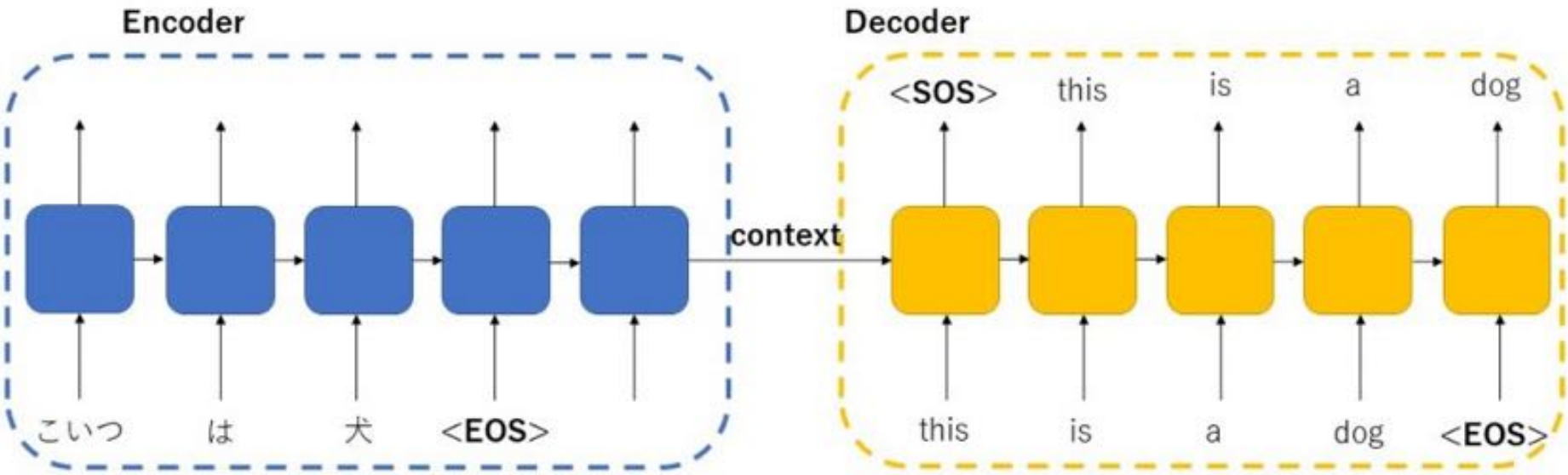
Séquence-to-séquence



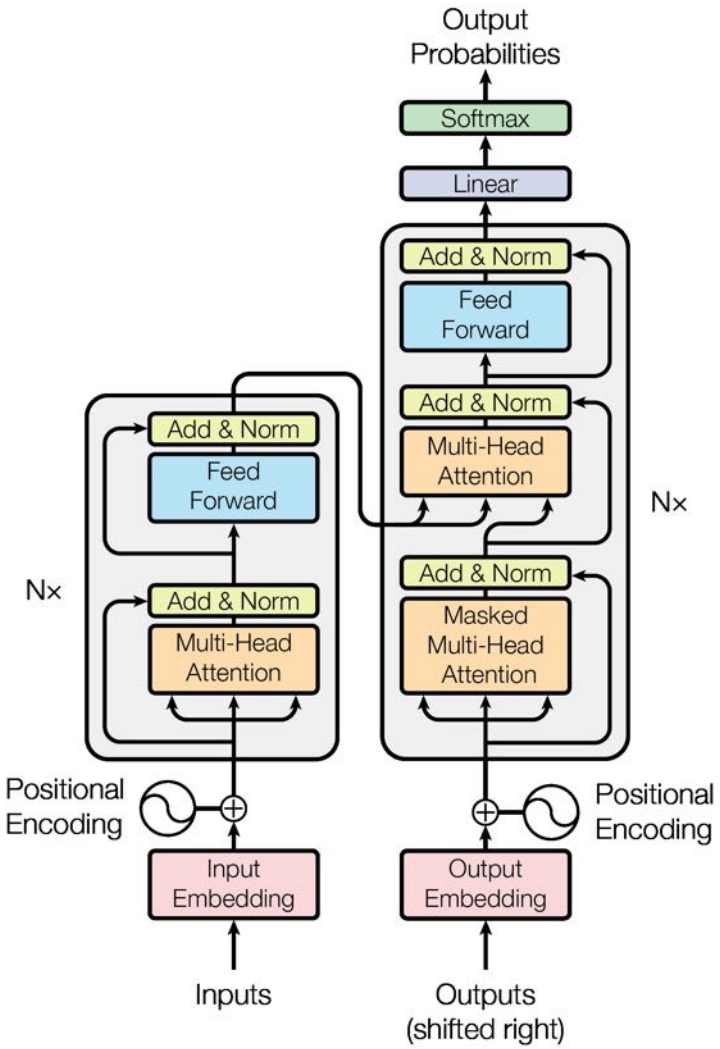
source: http://jalammar.github.io/illustrated-transformer/

Encodage du contexte

Source → Encodeur → CONTEXTE → Décodeur → Sorties



La transformation qui change la donne



Transformers

- Arrivée en 2017 : Vaswani et al, NIPS'17

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

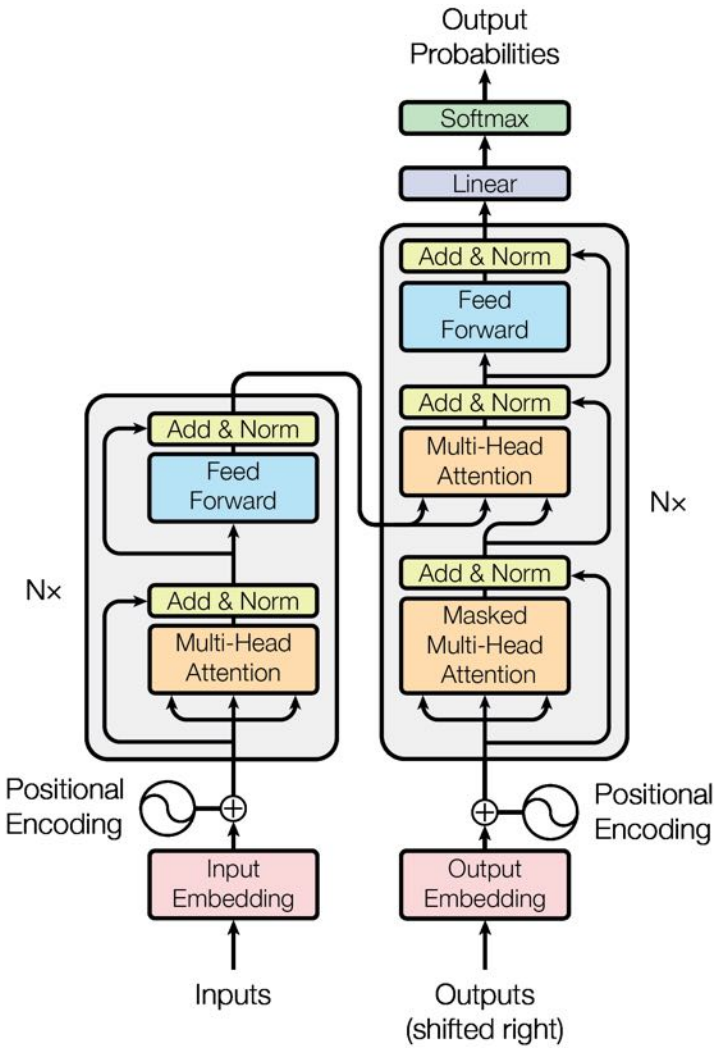
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the

- Séquence d'empilement de réseaux de transfo elf-attention (« auto-attention » ?)
- Encodeur et décodeur fonctionne de manière identique mais :
  - Encodeur calcule sa sortie en une fois, et transmets (comme contexte) au décodeur
  - Décodeur augmente ses entrées au fur et à mesure de l'inférence des sorties
    - Autorégressif : chaque sortie dépend des sorties précédentes

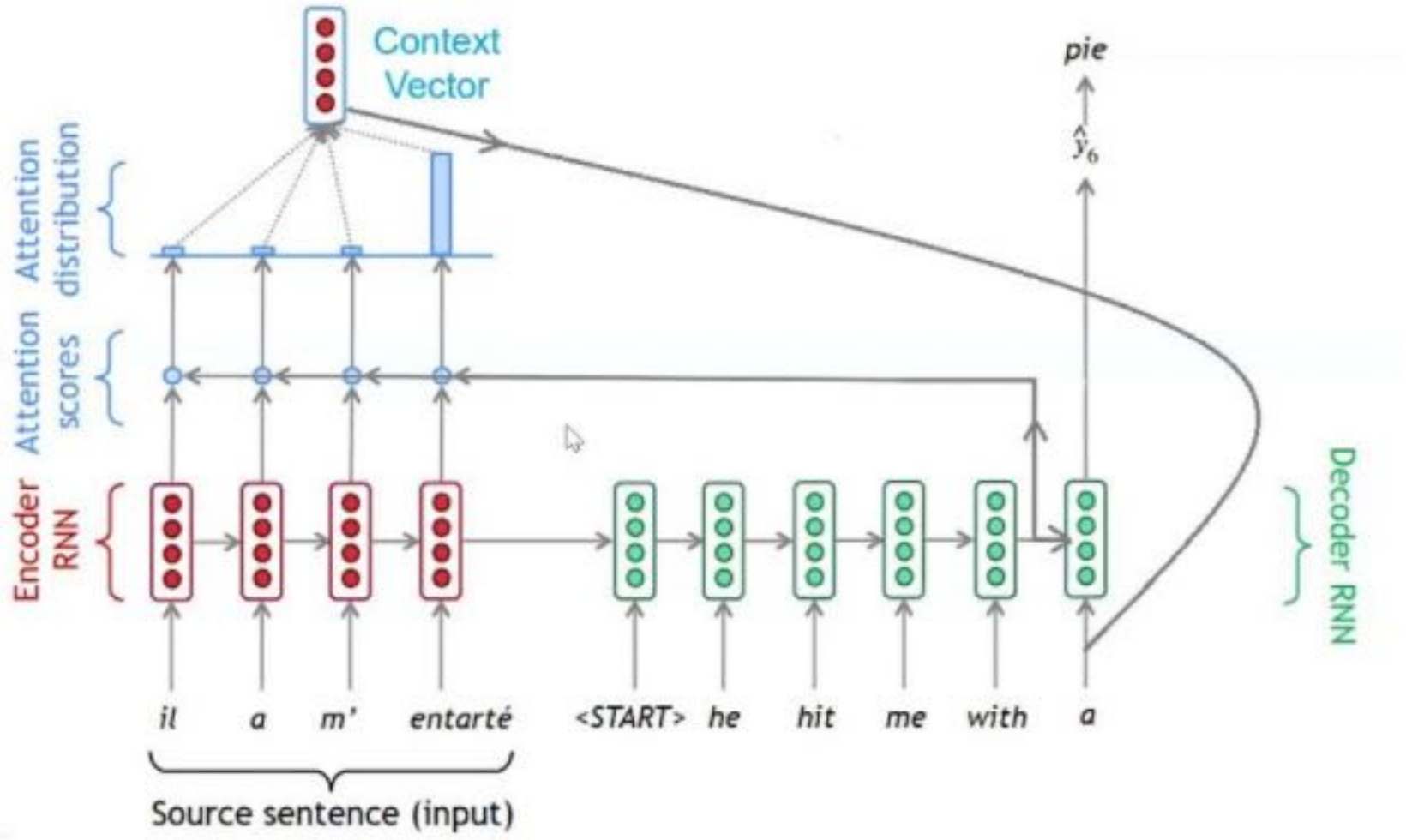
- ➔ Self attention

AVIGNON UNIVERSITÉ

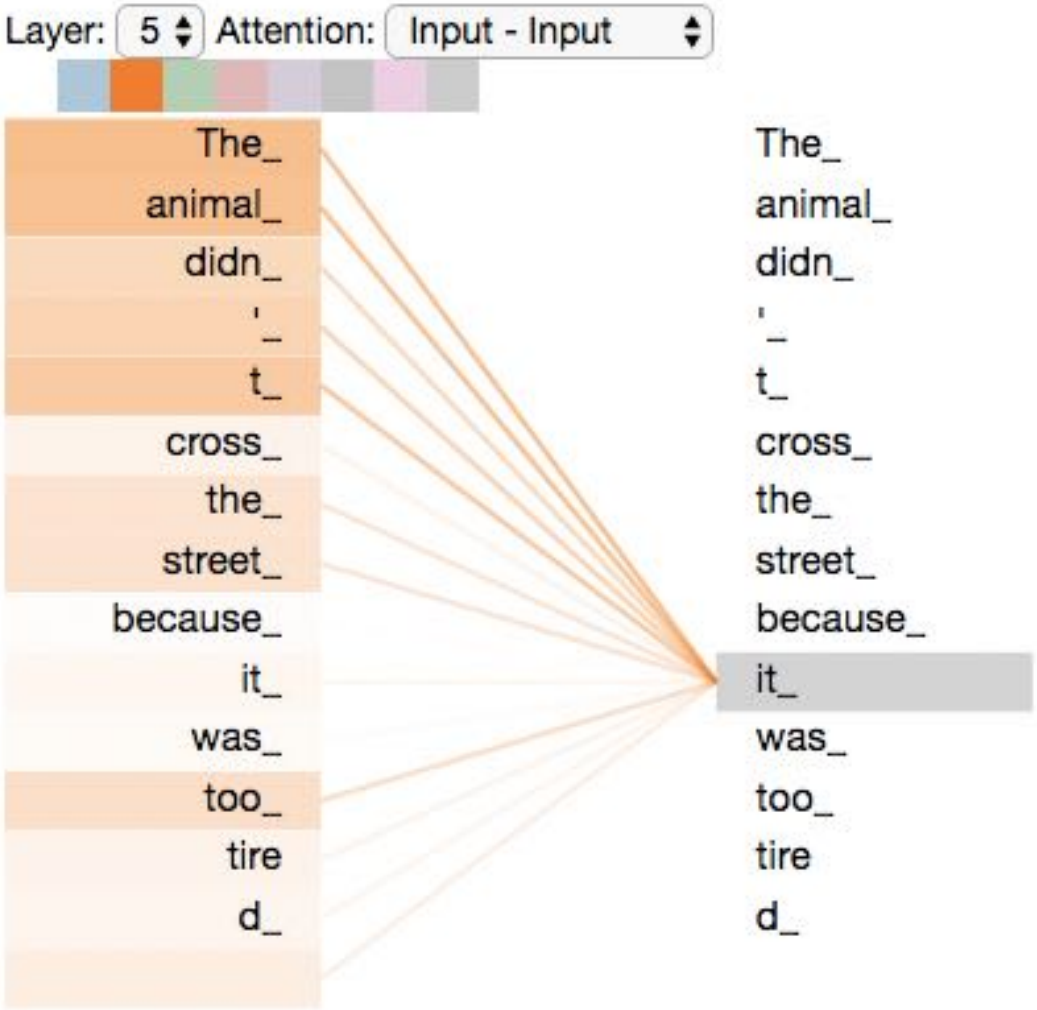
Transformer complet



Attention ? (tentative)

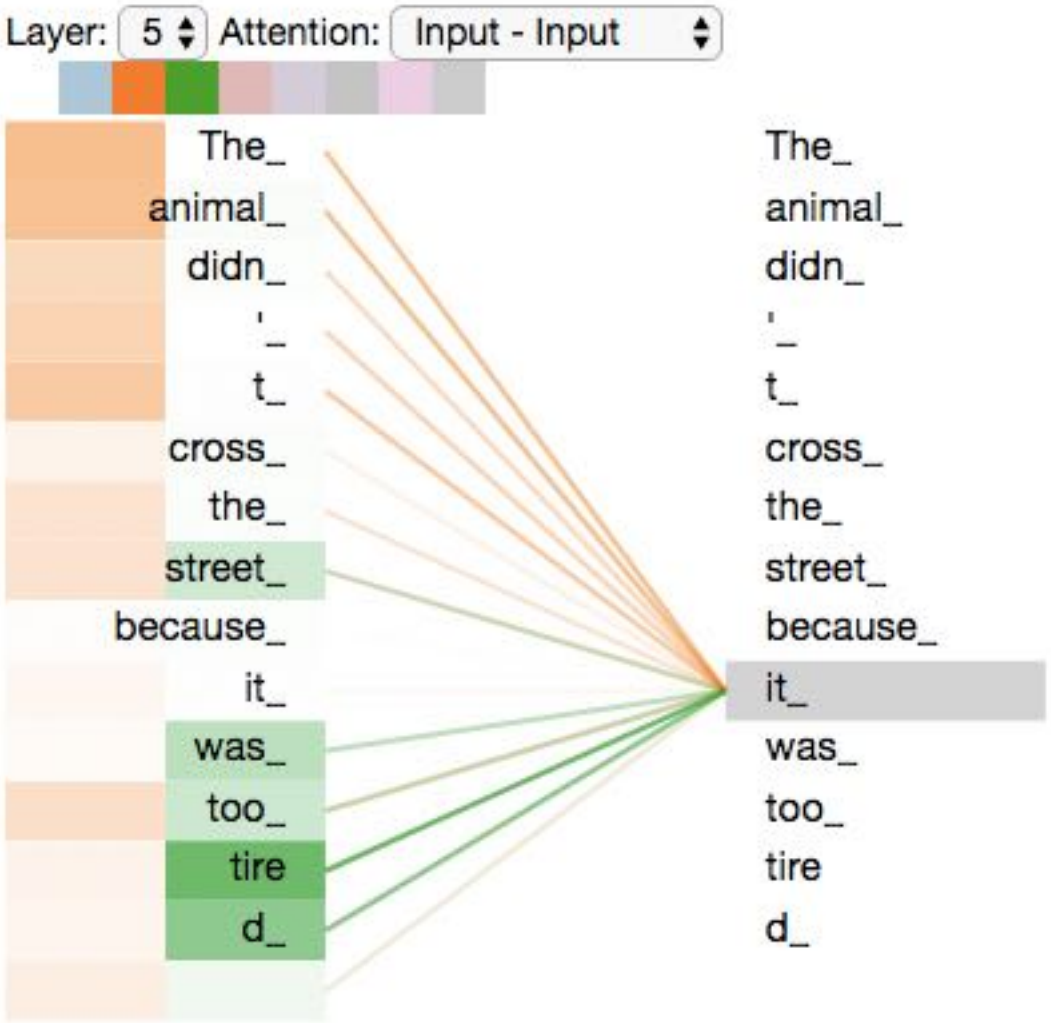


Multi-têtes de Self-attention



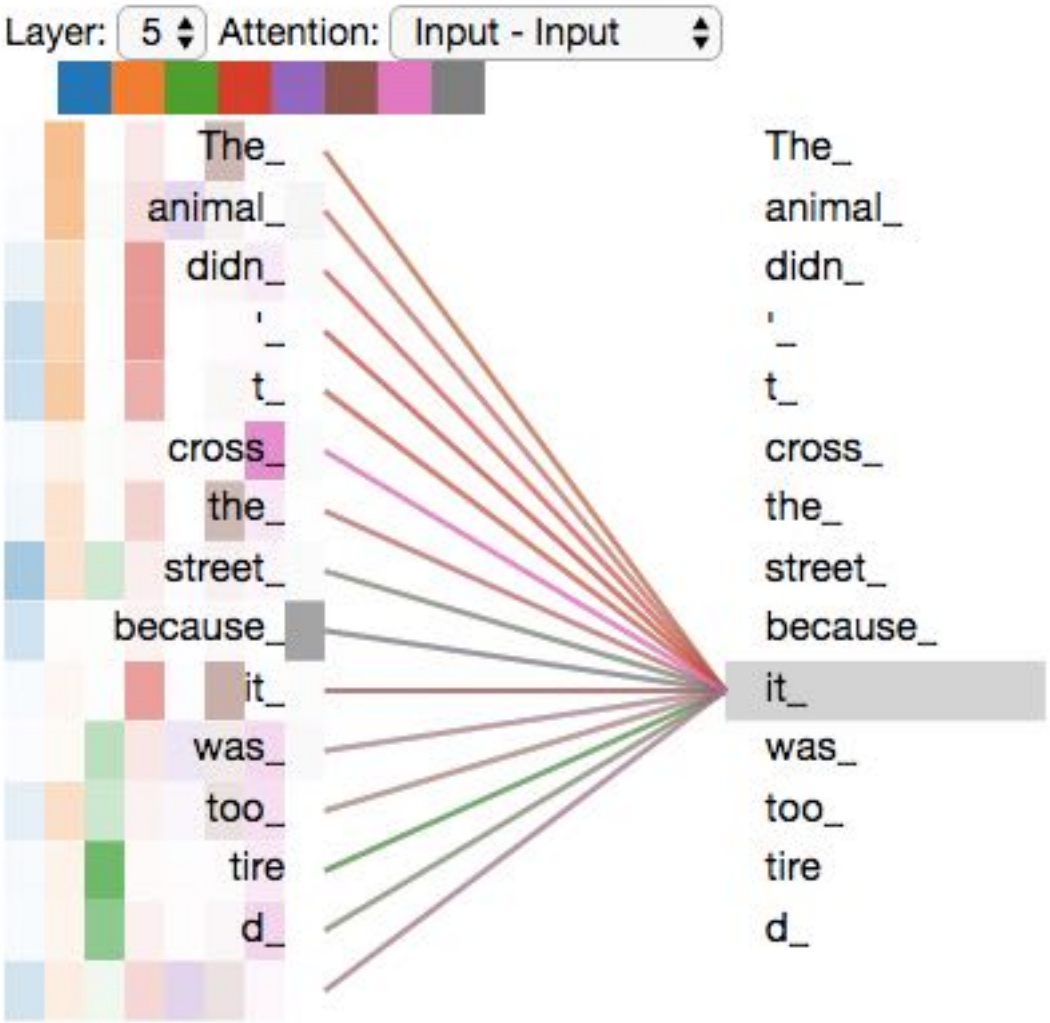
source: http://jalammar.github.io/illustrated-transformer/

Multi-têtes de Self-attention



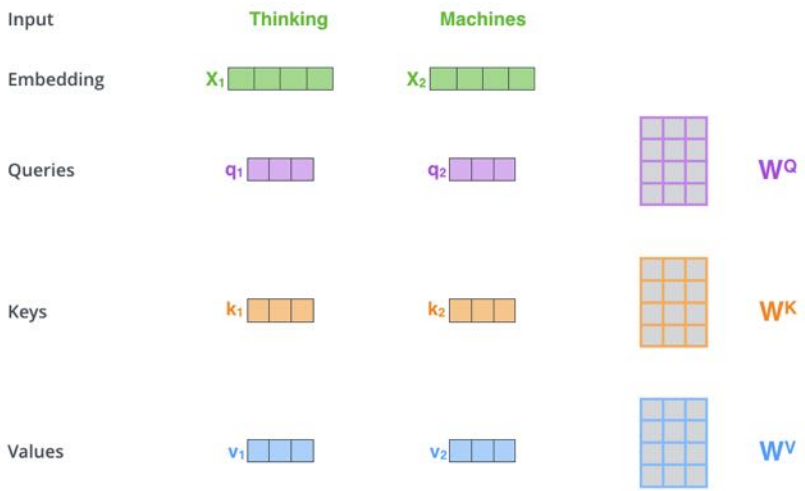
source: http://jalammar.github.io/illustrated-transformer/

Multi-têtes de Self-attention



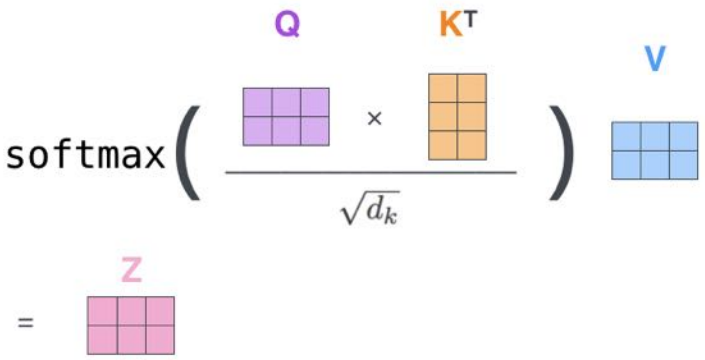
source: http://jalammar.github.io/illustrated-transformer/

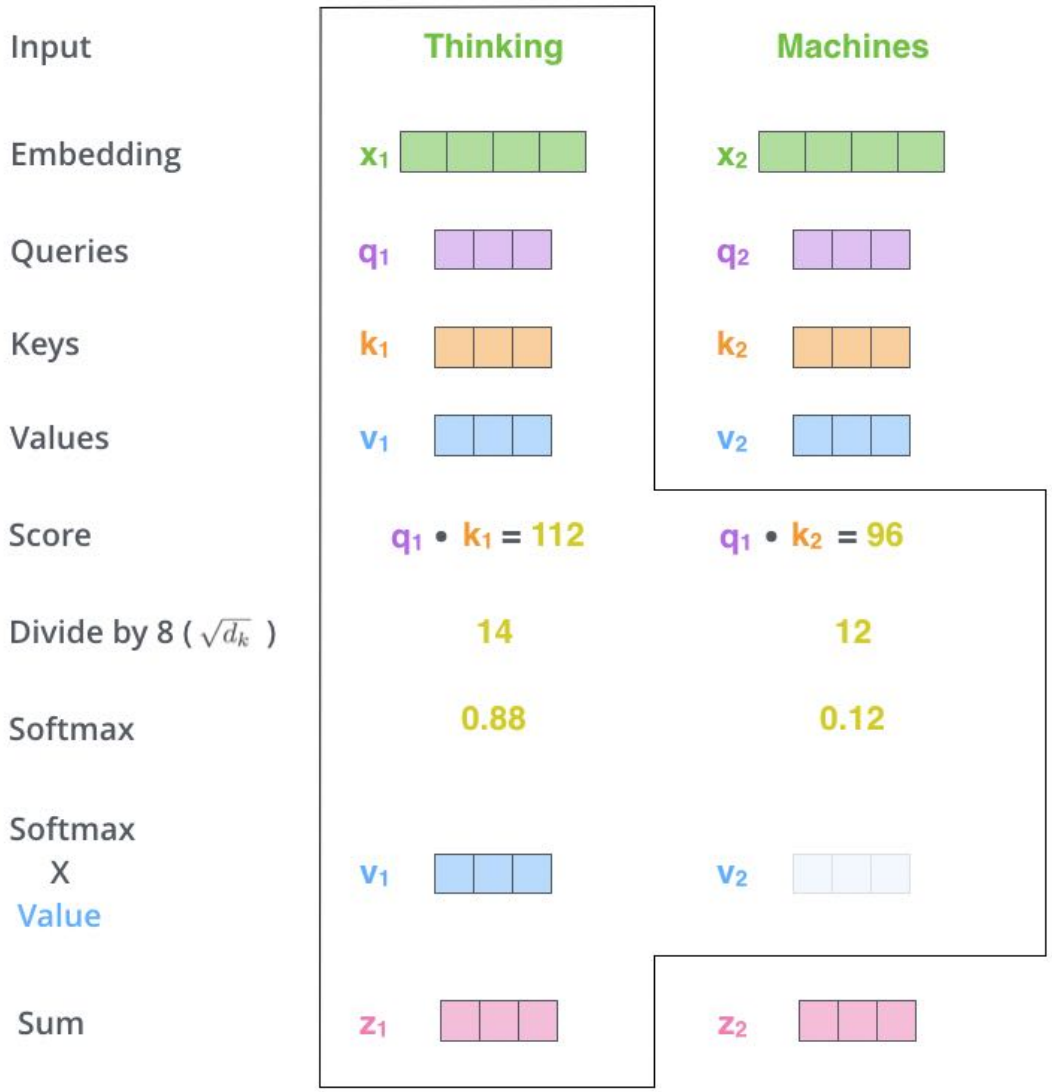
Calcul de la self-attention



source: http://jalammar.github.io/illustrated-transformer/

Calcul de la self-attention

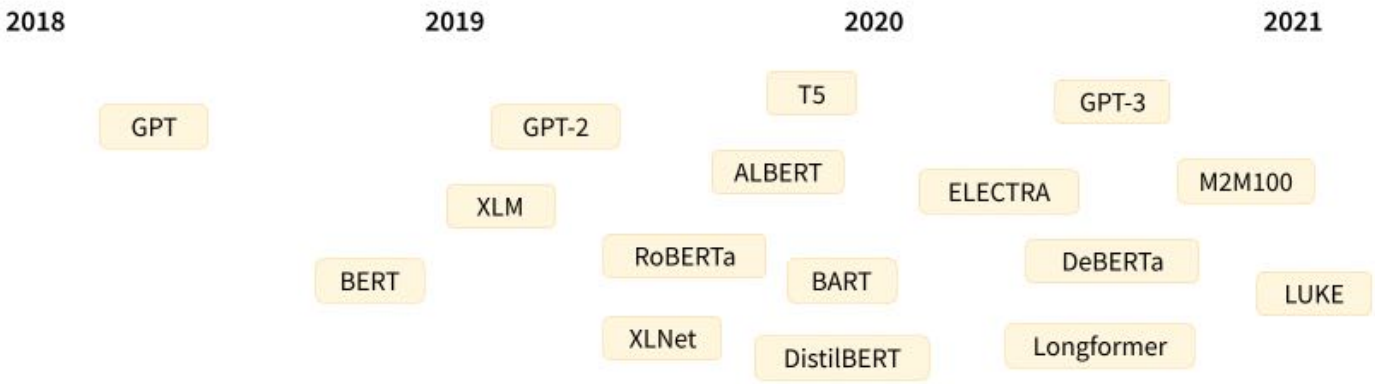




source: http://jalammar.github.io/illustrated-transformer/

La Transformers Family

- Dérivation de sous-Transformers qui vont encore accroître leur efficacité selon les cas d'usages
- Taxonomie :
  - Les encodeurs (*classifieurs*) : BERT, ElMO, RoBERTa, BART…
  - Les décodeurs (*générateurs*) : GPT-*n*, OPT, BLOOM…



La traduction, une tâche parmi d'autres

- Depuis la traduction usage des Transformers généralisé à toutes les tâches de manipulation de séquences
  - Notion de Large Language Models (LLMs, ou Pre-trained LM)
- Principalement en NLP :
  - Cas de GLUE : test ultime NLP (~20 tâches variées)
  - Conversations (Meta BlenderBot, Google LamDA, Sparrow…)
  - Q&A, search (ChatGPT, Bard…)
- Mais aussi pour la manipulation de séquences biologiques, par exemple

Cas de BERT

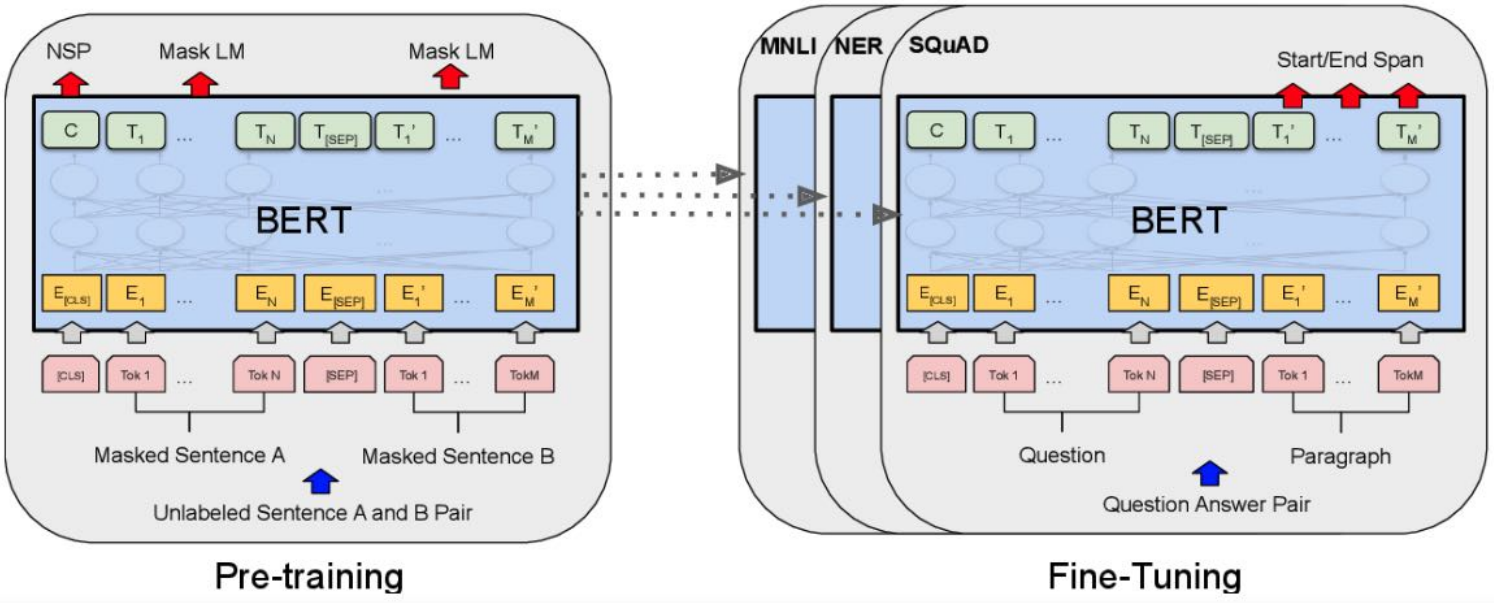


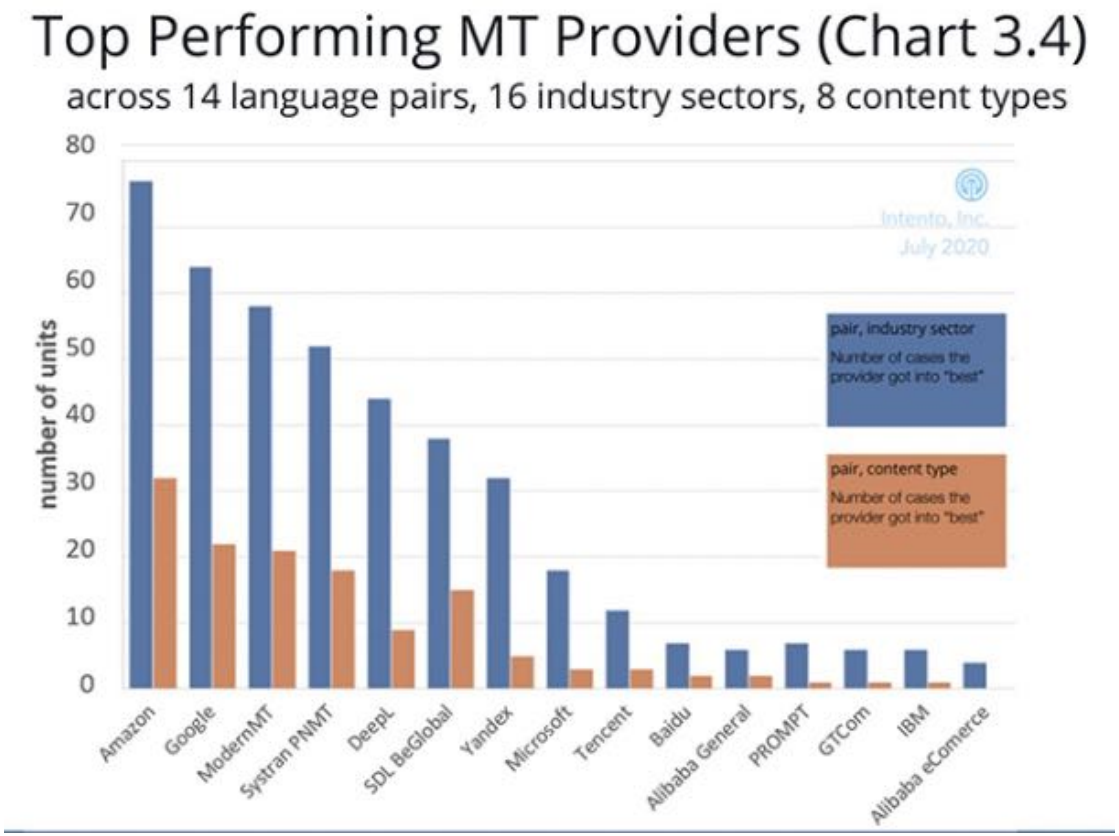
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers). Collapse

Published in North American Chapter of the Association for Computational Linguistics 2019

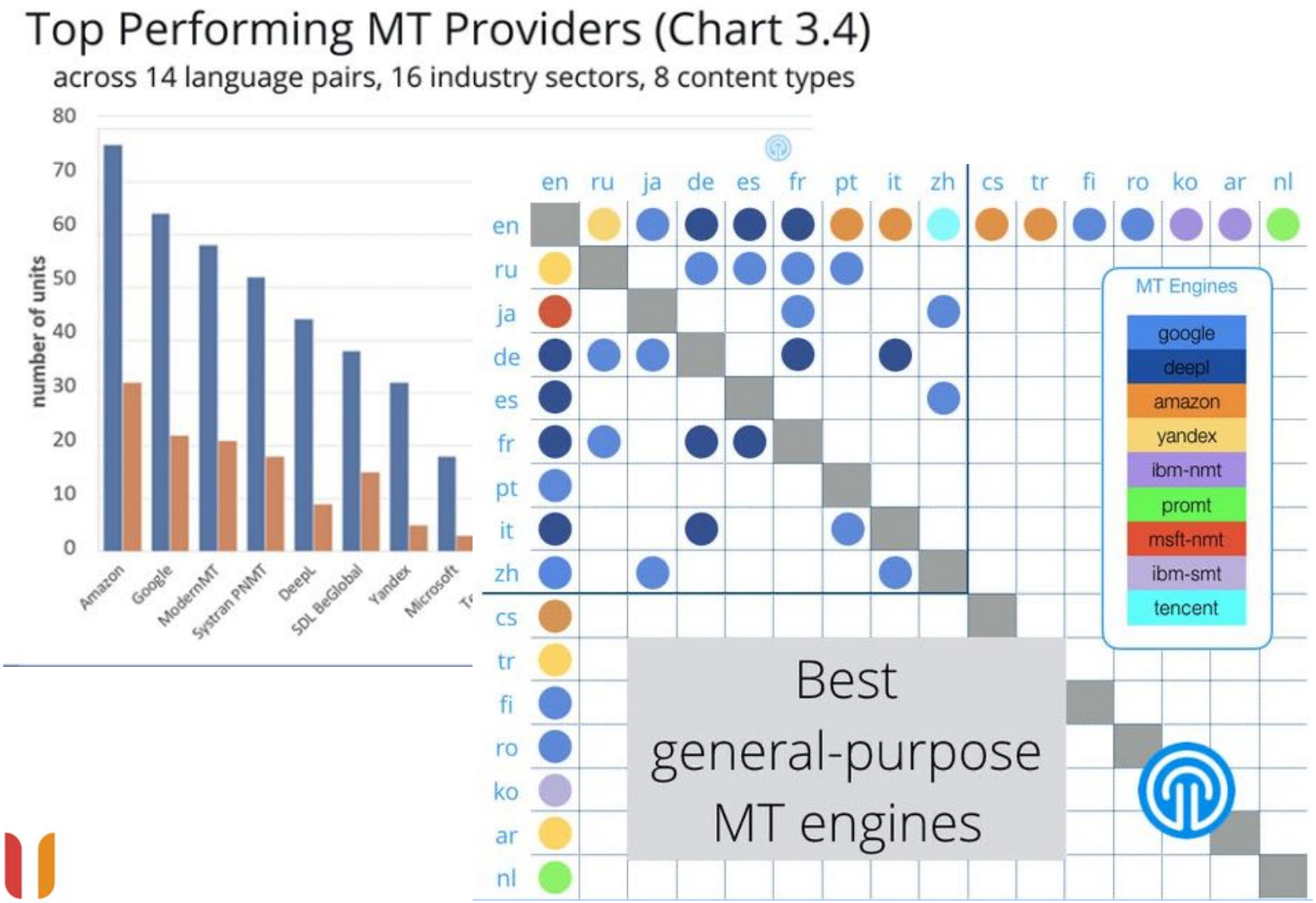
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

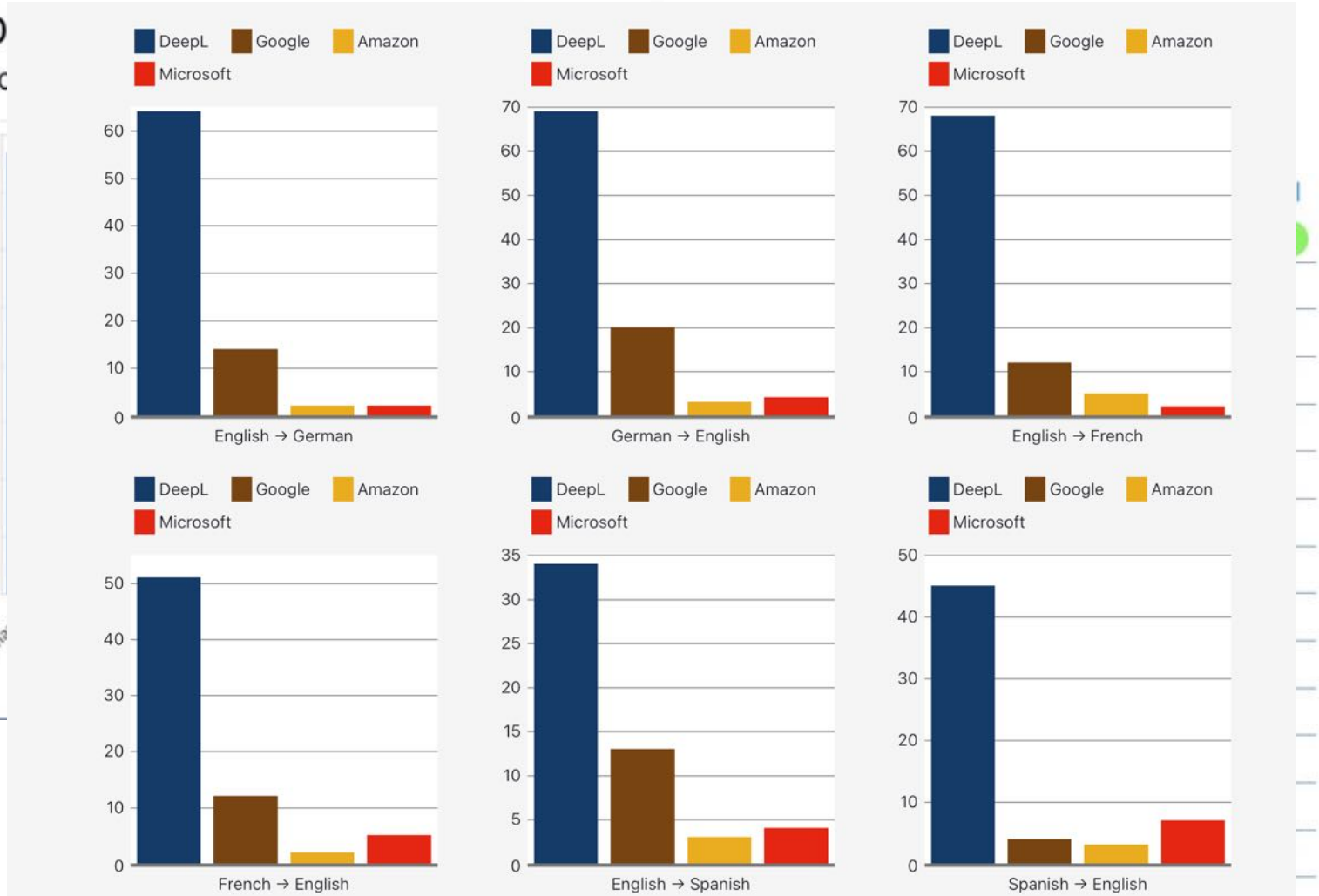
Performances



Performances



Performances



119 paragraphs from different domains were translated by the various systems. External professional translators were hired to review the translations - without information about which system produced which translation. The graph displays how often each system's translations were rated better than all other translations. Not shown are cases where several systems produced the best translation. The tests were performed in January 2020.

AVIGNON UNIVERSITÉ

Acteurs majeurs

- GAFAM, BATX… (surtout compagnies du search, mais pas que…)
  - Microsoft Translator
  - Google Translate
  - Amazon Translate
  - IBM Watson Language Translator
  - Yandex
  - Baidu
- Companies spécialisées
  - DeepL
  - Systran
  - TransPerfect
  - RWS
- Myriade de start-ups (capitalisant sur la ressource LLM)
- Une part du marché IT toujours grandissante

AVIGNON UNIVERSITÉ

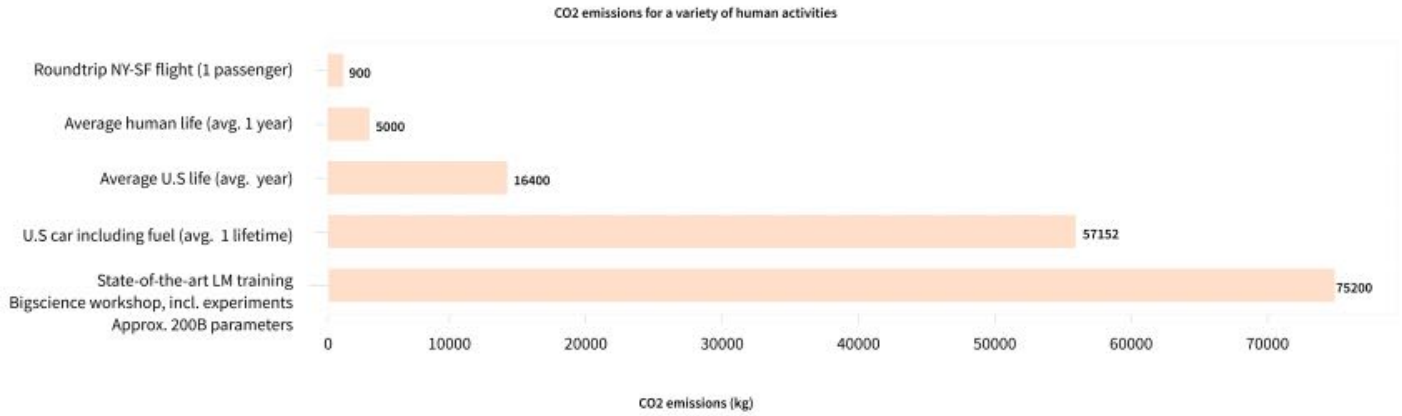
Limites de l'approche

- Modèles très lourds, chers à apprendre (millions $), complexe à utiliser (cloud GPU)
  - Distillation : mécanisme d'apprentissage permettant de ne conserver que les paramètres utiles à une tâche en particulier (pas de retour en arrière)
- Adaptation pas aisée
  - Adaptation : fine-tuning, adapters…
- Robustesse encore faible
  - Variations dues aux erreurs d'entrée
  - Due à la non-prise en compte du sens

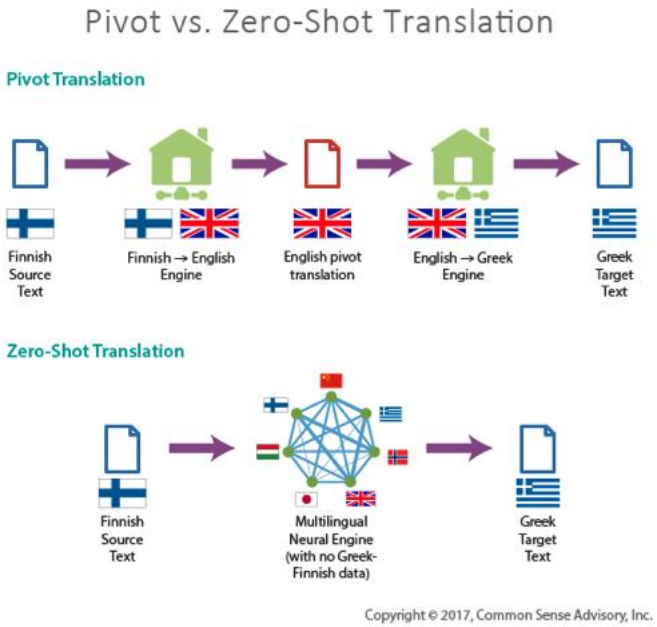
AVIGNON UNIVERSITÉ

Limites de l'approche

- Modèles très lourds, chers à apprendre (millions $), complexe à utiliser (cloud GPU)
  - Distillation : mécanisme d'apprentissage permettant de ne conserver que les paramètres utiles à une tâche en particulier (pas de retour en arrière)
- Ada
  -
- Rob
  -
  - Due à la non-prise en compte du sens



Enjeux…

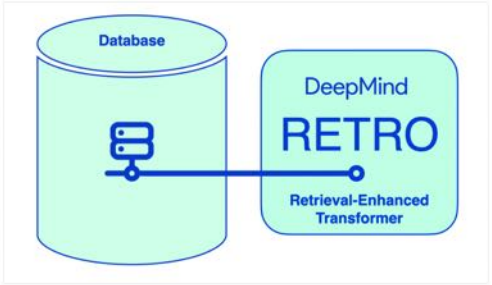


- Démocratisation des systèmes de traduction
  - Accès aux LLMs (cas du libre BLOOM)
  - Réduction des coûts de développement
- Systèmes multilingues
  - Configuration langue pivot vs multi-langues (BLOOM : 46 langues !)
- Langues peu dotées
  - Les résultats majeurs portent sur 5 à 10 langues, parmi les plus répandues (sur ~7500 langues parlées dans le monde)
  - Coût possibilité de mise en œuvre sur des langues peu ressourcées (problème des langues orales…)

AVIGNON UNIVERSITÉ

Enjeux

- Spécialisation
  - Mémoires de traduction
  - Domaines spécialisés



RETRO incorporates information retrieved from a database to free its parameters from being an expensive store of facts and world knowledge.

- Amélioration continue (*continual learning*)
  - Nouveaux termes, nouvelles formulations

- Droits sur les traductions
  - L'origine des données d'apprentissage (qui peuvent être libres de droits mais pas forcément) pose le problème de la propriété des résultats

- Et plein d'autres

AVIGNON UNIVERSITÉ