

Traitement Automatique des Langues Naturelles

Exemple d'applications: la traduction automatique

Chloé Braud, Philippe Muller

Master IAFA 2024-2025

Exemple d'application en TAL: Traduction automatique

- La Traduction automatique est une tâche difficile
- Test ultime pour la compréhension de la langue
- Pour correctement traduire il faut
 - comprendre la langue **source**
 - comprendre la langue **cible**
 - avoir des connaissances sur le **thème** du texte à traduire
 - avoir des connaissances sur la **culture**, valeurs, traditions, attentes des locuteurs dans les deux langues

nation, the same geographical area, the same cultural tradition: e.g., the Flemish, French, and German communities of Belgium; a Bantu language; the English language; etc.). 1. the action of translating from one language into another; a new rendering of something in another language. 2. a version of something resulting from such a process. 3. a new translation of Plato. 4. a new version of a work of art or literature, e.g., a film or painting.



Une longue histoire

- Warren Weaver (1949):

"J'ai sous les yeux un texte écrit en russe, mais je vais prétendre qu'il est vraiment écrit en anglais et qu'il a été codé avec des symboles étranges. Tout ce que je dois faire est d'enlever le code afin de récupérer les informations contenues dans le texte."

- Première démonstration par IBM en 1954 avec un système de base (traduction mot par mot)
- Mais la traduction automatique s'est avérée beaucoup plus dure que prévu



Intérêt politique/commercial

- L'UE dépense plus d'un milliard d'euros en frais de traduction chaque année ; la (semi-)automatisation permettrait d'économiser beaucoup d'argent
- Les États-Unis ont beaucoup investi dans la traduction pour faciliter l'acquisition de renseignements
- Google Translate est utilisé quotidiennement par plus de 500 millions d'utilisateurs et traite plus d'un milliard de traductions chaque jour



Intérêt académique

- informatique, linguistique, statistique, intelligence artificielle, sciences cognitives, ...
- Traduction automatique est “IA-difficile” : nécessite une solution au problème d’IA général de représenter et de raisonner sur divers types de connaissances (linguistique, du monde, ...)
- ... bien que tous les systèmes restent jusqu’à présent assez superficiels

$$\int (x^n + a^n)^r dx = \int (x^n + a^n)^{r-1} - a \int \frac{dx}{x^n + a^n}$$
$$\int \frac{dx}{x^n(x^n + a^n)} = \frac{1}{a^n} \int \frac{dx}{x^{n-n}(x^n + a^n)^{r-1}} - \frac{1}{a^n} \int \frac{dx}{x^n}$$
$$\int \frac{dx}{x\sqrt{x^n + a^n}} = \frac{1}{n\sqrt{a^n}} (\ln \int \frac{\sqrt{x^n + a^n} - \sqrt{a^n}}{\sqrt{x^n + a^n} + \sqrt{a^n}} =$$
$$\int \frac{dx}{x^{m-n}(x^n - a^n)^r} = \frac{1}{a^n} \int \frac{dx}{x^{m-n}(x^n - a^n)^{r-1}} - \frac{1}{a^n}$$
$$= \int \frac{dx}{x\sqrt{x^n - a^n}} = \frac{2}{n\sqrt{a^n}} \cos^{-1} \sqrt{\frac{a^n}{x^n}} =$$
$$\int \frac{dx}{x(x^n + a^n)} = \frac{1}{n a^n} (\ln \frac{x^n}{x^n + a^n} + \int \frac{x^{n-1} dx}{x^n + a^n})$$



ordre des mots différent

- ordre des mots en anglais est
sujet-verbe-objet
ordre des mots en japonais est
sujet-objet-verbe
- anglais: *Merkel visited Trump*
japonais: *Merkel Trump visited*
- anglais: *Reporters said Merkel visited Trump*
japonais: *Reporters Merkel Trump visited said*



ordre des mots différent

- ordre des mots en anglais est *adjectif-nom*
ordre des mots en français est *nom-adjectif*
- anglais: *a red house*
français: *une maison rouge*
- anglais: *almost double the usual oatmeal biscuits*
français: *presque le double des biscuits à l'avoine habituels*



ambiguïté du sens des mots

- Un mot peut avoir différents sens
- e.g. *avocat*

ambiguïté du sens des mots

- Un mot peut avoir différents sens
- e.g. *avocat*



ambiguïté du sens des mots

- Un mot peut avoir différents sens
- e.g. *avocat*



ambiguïté du sens des mots

- Un mot peut avoir différents sens
- e.g. *avocat*



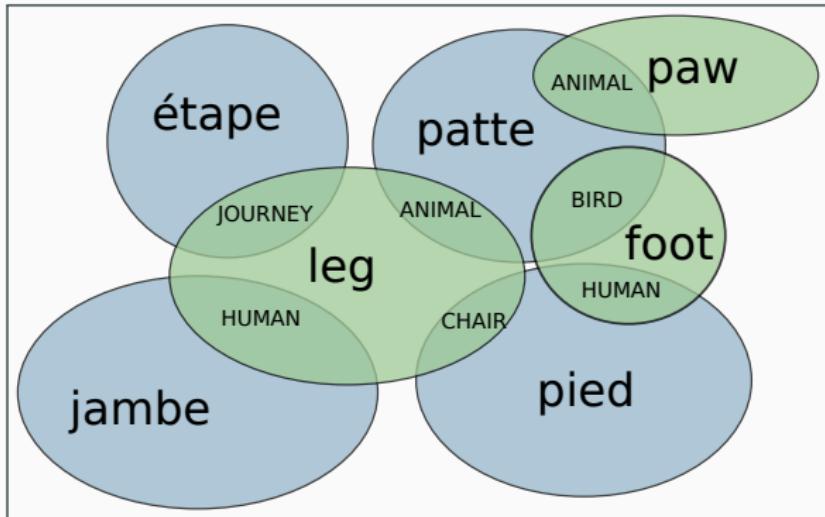
avocado



lawyer

Difficultés

ambiguïté du sens des mots



Difficultés

pronoms

- Certains langues comme l'espagnol peuvent omettre les pronoms sujets; langues “pro-drop”
- En espagnol, la flexion verbale indique souvent quel pronom doit être restauré (mais pas toujours)
 - -o = je
 - -as = tu
 - -a = il/elle
 - -amos = nous
 - -an = ils/elles
- Quand le système doit-il utiliser *il, elle ; ils, elles?*
- Quand le système doit-il utiliser *he, she, it?*



temps du verbe

- français: *je bois du lait*



temps du verbe

- français: *je bois du lait*
- anglais: *I drink milk ?*



temps du verbe

- français: *je bois du lait*
- anglais: *I drink milk ?*
- anglais: *I'm drinking milk ?*



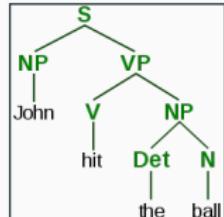
expressions idiomatiques

- *pomme de discorde*
- *casser sa pipe*
- *vendre la peau de l'ours avant de l'avoir tué*



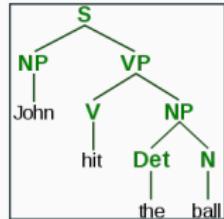
Différentes approches

- Traduction automatique **symbolique**
 - basée sur la linguistique et la logique
- Traduction automatique **statistique**
 - basée sur le théorème de Bayes
- Traduction automatique **neuronale**
 - basée sur des réseaux de neurones



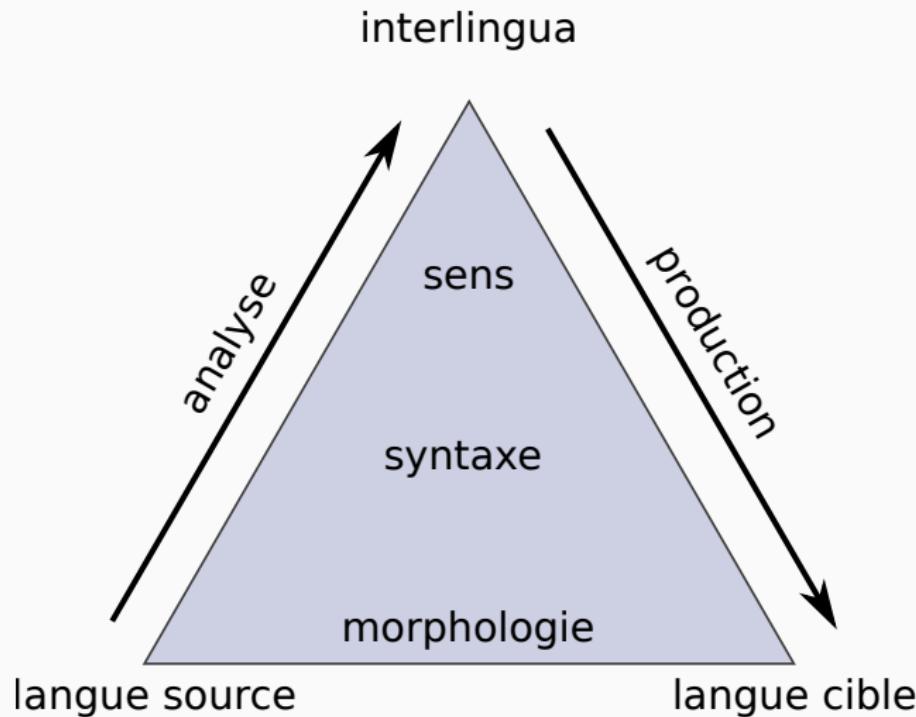
Différentes approches

- Traduction automatique symbolique
 - basée sur la linguistique et la logique
- Traduction automatique statistique
 - basée sur le théorème de Bayes
- Traduction automatique neuronale
 - basée sur des réseaux de neurones



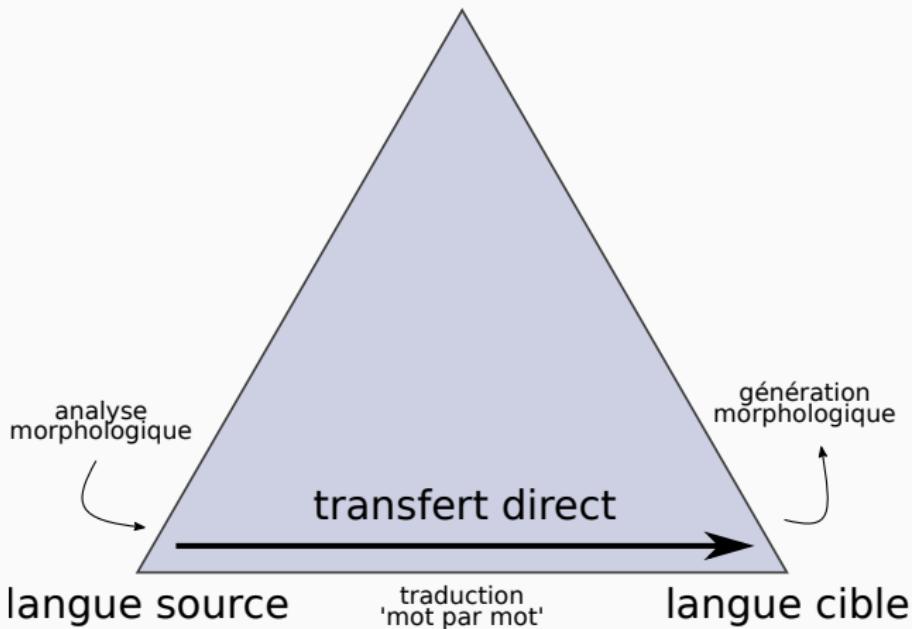
Traduction symbolique

Triangle de Vauquois



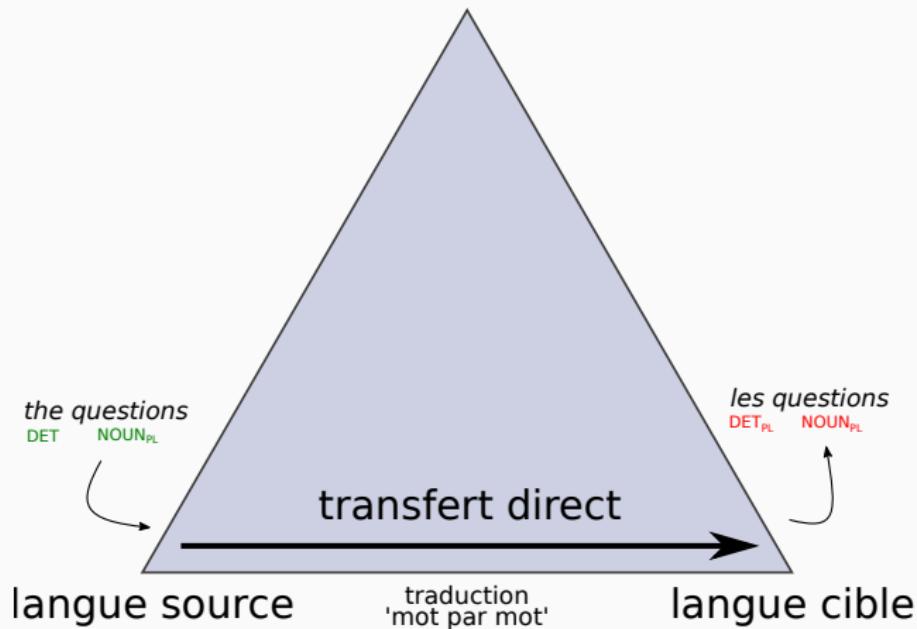
Traduction symbolique

Transfert direct



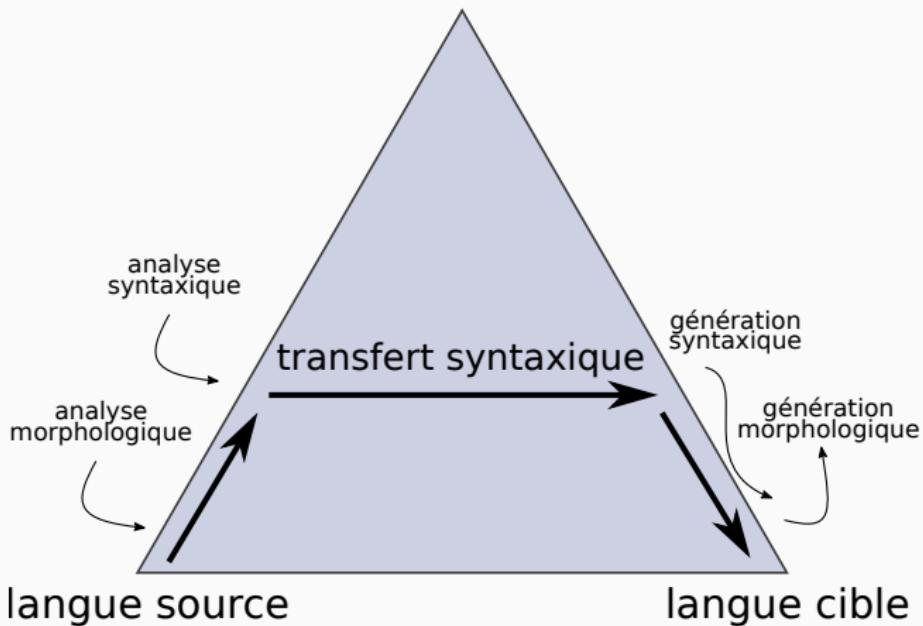
Traduction symbolique

Transfert direct



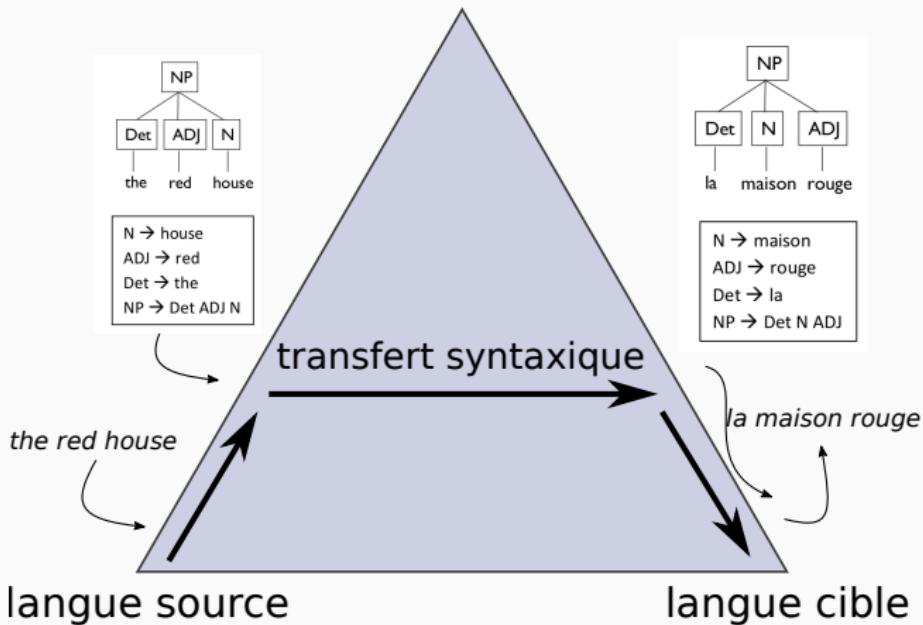
Traduction symbolique

Transfert syntaxique



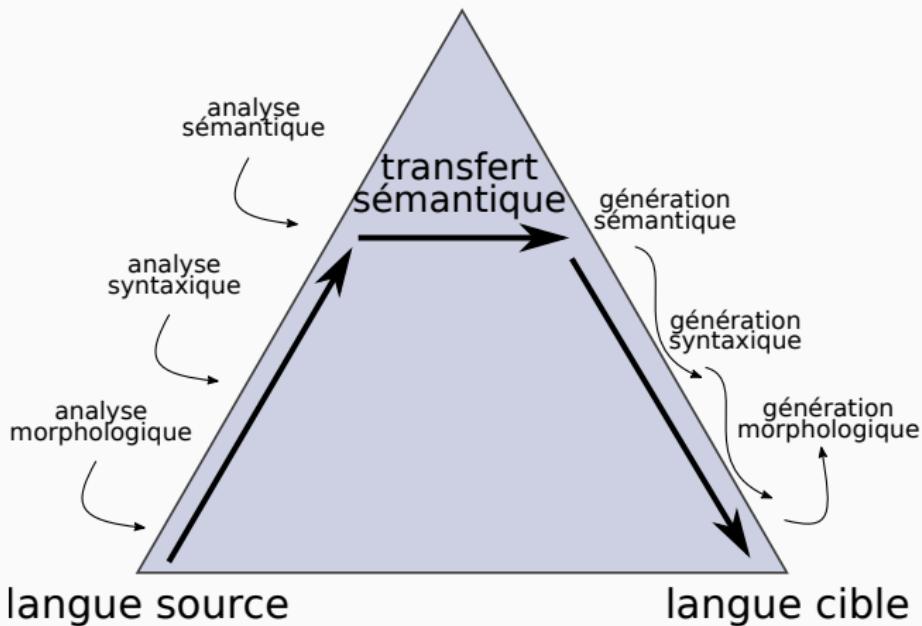
Traduction symbolique

Transfert syntaxique



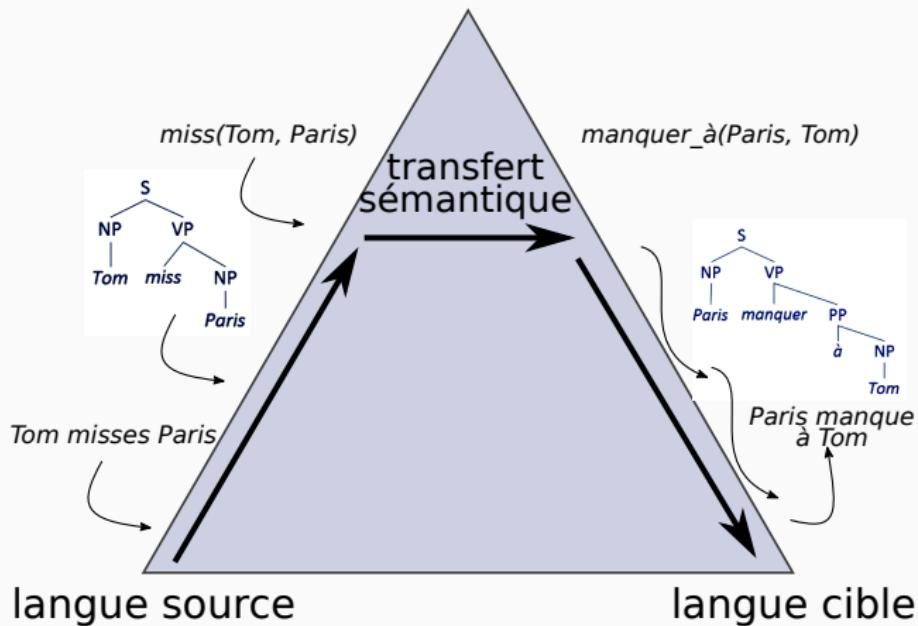
Traduction symbolique

Transfert sémantique



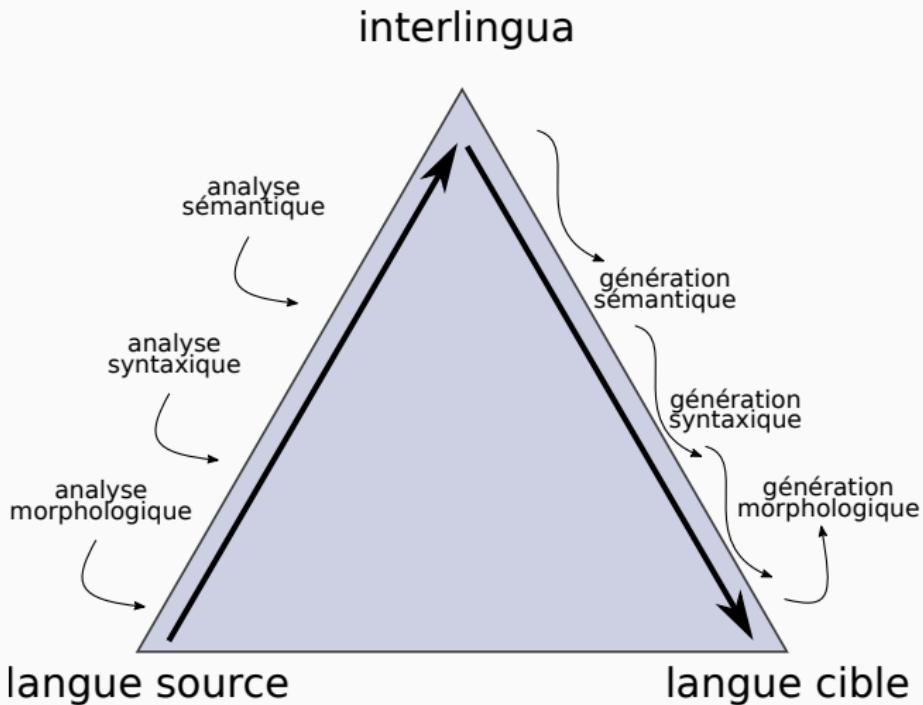
Traduction symbolique

Transfert sémantique



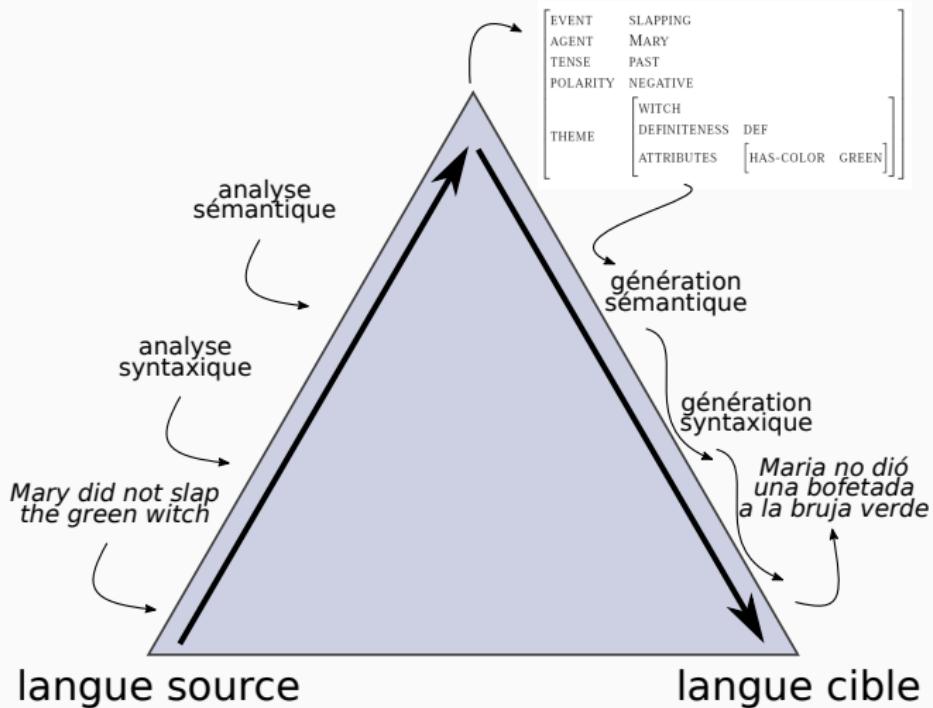
Traduction symbolique

Interlingua



Traduction symbolique

Interlingua



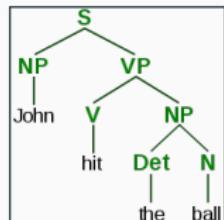
Inconvénients

- Traduction symbolique marche bien pour des applications très restreintes
- Difficultés de généralisation pour domaine ouvert (ambiguïtés, exceptions, ...)
- Création des règles se fait manuellement
- Développement coûteux et laborieux



Différentes approches

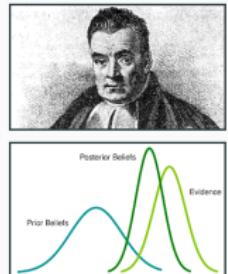
- Traduction automatique symbolique
 - basée sur la linguistique et la logique
- **Traduction automatique statistique**
 - basée sur le théorème de Bayes
- Traduction automatique neuronale
 - basée sur des réseaux de neurones



Probabilités

- Trouver la phrase anglaise la plus probable étant donnée une phrase en langue française

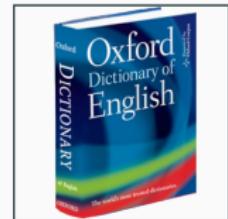
$$\begin{aligned}\hat{e} &=_{e \in \text{anglais}} p(e | f) \\ &=_{e \in \text{anglais}} \frac{p(f | e)p(e)}{p(f)} \\ &=_{e \in \text{anglais}} p(f | e)p(e)\end{aligned}$$



Traduction statistique

Pourquoi Bayes ?

- Pourquoi ne modélise-t-on pas directement $p(e | f)$?
- En utilisant $p(f | e)P(e)$, on est capable de diviser la tâche
 - $p(e)$ s'occupe du bon usage de l'anglais
 - $p(f | e)$ veille à ce que le français correspond à l'anglais
 - On peut les entraîner indépendamment



Traduction statistique

Example

Nous regardons la télévision

	bon usage d' anglais? $P(E)$	bonne correspondance au français? $P(F E)$
We look the television	✗	✓
It bike nothing bottle	✗	✗
In television we watch	✗	✓
That seems fairly reasonable	✓	✗
We are watching television	✓	✓
Are watching television we	✗	✓
He felt relieved	✓	✗

Traduction statistique

Example

Nous regardons la télévision

	bon usage d' anglais? $P(E)$	bonne correspondance au français? $P(F E)$
We look the television	✗	✓
It bike nothing bottle	✗	✗
In television we watch	✗	✓
That seems fairly reasonable	✓	✗
We are watching television	✓	✓
Are watching television we	✗	✓
He felt relieved	✓	✗

Traduction statistique

$p(e)$ – bon usage

- Comment calculer la probabilité d'une phrase,
e.g. *le chien courait à travers la pelouse* ?
- On pourrait compter le nombre de fois qu'on voit la phrase, et diviser par le nombre total de phrases dans notre corpus

$$\frac{\text{compte}(\text{le chien courait travers la pelouse})}{\text{nombre total de phrases}}$$



Traduction statistique

$p(e)$ – bon usage

- Comment calculer la probabilité d'une phrase,
e.g. *le chien courait à travers la pelouse* ?
- On pourrait compter le nombre de fois qu'on voit la phrase, et diviser par le nombre total de phrases dans notre corpus

$$\frac{\text{compte}(\textit{le chien courait travers la pelouse})}{\text{nombre total de phrases}} = \frac{0}{10^{14}} = 0$$



modèle de langue

Comment calculer $p(\text{le}, \text{chien}, \text{courait}, \text{à}, \text{travers}, \text{la}, \text{pelouse})$?

- On calcule la probabilité jointe par décomposition en probabilités conditionnelles
- $p(\text{le}) \times p(\text{chien}|\text{le})$
 $\times p(\text{courait}|\text{le chien})$
 $\times p(\text{à}|\text{le chien courait})$
 $\times p(\text{travers}|\text{le chien courait à})$
 $\times p(\text{la}|\text{le chien courait à travers})$
 $\times p(\text{pelouse}|\text{le chien courait à travers la})$



cf cours sur les n-grammes/modèles de langue

modèle de langue

Comment calculer $p(\text{le}, \text{chien}, \text{courait}, \text{à}, \text{travers}, \text{la}, \text{pelouse})$?

- On calcule la probabilité jointe par décomposition en probabilités conditionnelles
- $p(\text{le}) \times p(\text{chien}|\text{le})$
 $\times p(\text{courait}|\text{le chien})$
 $\times p(\text{à}|\text{le chien courait})$
 $\times p(\text{travers}|\text{le chien courait à})$
 $\times p(\text{la}|\text{le chien courait à travers})$
 $\times p(\text{pelouse}|\text{le chien courait à travers la})$
- ... en utilisant un contexte limité
cf cours sur les n-grammes/modèles de langue



modèle de langue

Comment calculer $p(\text{le}, \text{chien}, \text{courait}, \text{à}, \text{travers}, \text{la}, \text{pelouse})$?

- On calcule la probabilité jointe par décomposition en probabilités conditionnelles
- $p(\text{le}) \times p(\text{chien}|\text{le})$
 $\times p(\text{courait}|\text{le chien})$
 $\times p(\text{à}|\text{chien courait})$
 $\times p(\text{travers}|\text{courait à})$
 $\times p(\text{la}|\text{à travers})$
 $\times p(\text{pelouse}|\text{travers la})$
- ... en utilisant un contexte limité
cf cours sur les n-grammes/modèles de langue



Pierre de Rosette

- Pendant des siècles, la langue égyptienne était un mystère
- En 1799, une pierre avec un texte égyptien et sa traduction en grec était trouvée
- La tablette a permis de pouvoir traduire les hiéroglyphes



Traduction statistique

$p(f \mid e)$ – bonne correspondance

- Nos algorithmes utiliseront le même processus
- On utilise une grande collection de texte en deux langues: **corpus parallèle**
- Pierre de Rosette d'aujourd'hui :
 - actes du Parlement européen
 - sous-titres de films
 - sites web multilingues
 - ...



Traduction statistique

$p(f | e)$ – bonne correspondance

- On connaît les alignements entre phrase source et phrase cible
- On ne connaît pas les alignements des mots individuels
- Problème: comment proprement généraliser à partir d'exemples
- Il faut qu'on obtienne de bonnes statistiques pour construire le modèle $P(F | E)$



Intuition

Quel est le mot pour *ciel* en vietnamien ?

anglais In the beginning God created the **heavens** and the earth.

vietnamien Ban dây Đức Chúa Trời dung nên trời đất.

anglais God called the expanse **heaven**.

vietnamien Đức Chúa Trời đặt tên khoang không là trời.

anglais ... you are this day like the stars of **heaven** in number.

vietnamien ... các người đông nhu sao trên trời.

Intuition

Quel est le mot pour *ciel* en vietnamien ?

anglais In the beginning God created the **heavens** and the earth.

vietnamien Ban dây Đức Chúa Trời dung nên **trời** dây.

anglais God called the expanse **heaven**.

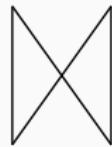
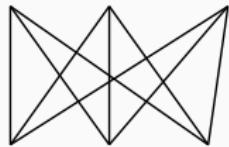
vietnamien Đức Chúa Trời đặt tên khoang không là **trời**.

anglais ... you are this day like the stars of **heaven** in number.

vietnamien ... các người đông nhu sao trên **trời**.

Alignment

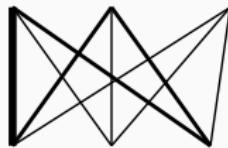
... la maison ... la maison blue ... la fleur ...



... the house ... the blue house ... the flower ...

Alignment

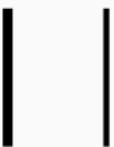
... la maison ... la maison blue ... la fleur ...



... the house ... the blue house ... the flower ...

Alignment

... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...

Alignment

... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...

Traduction statistique

Alignment

... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...



$$p(\text{la} | \text{the}) = 0.453$$

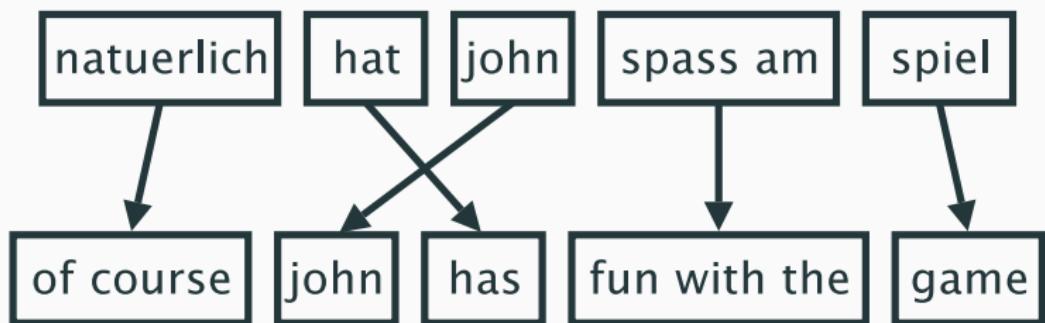
$$p(\text{le} | \text{the}) = 0.334$$

$$p(\text{maison} | \text{house}) = 0.876$$

$$p(\text{bleu} | \text{blue}) = 0.563$$

...

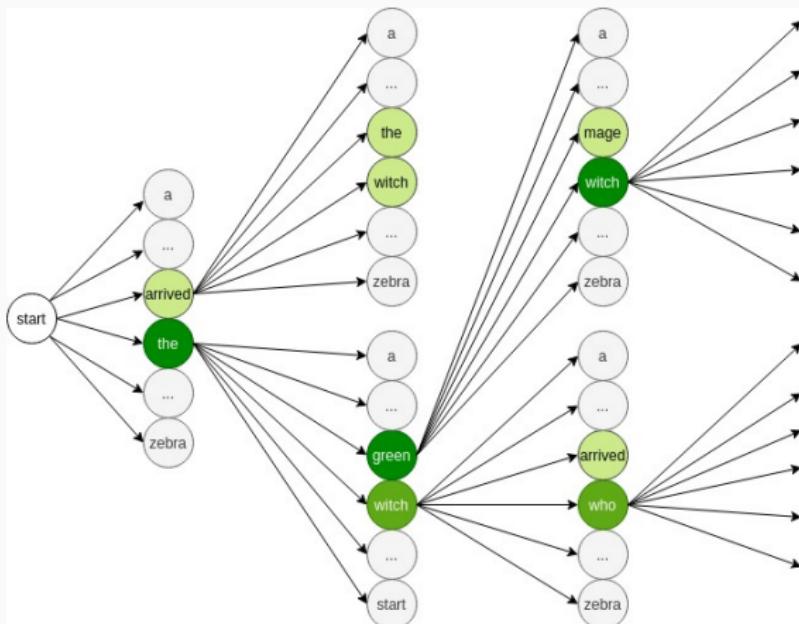
Alignment



- Liens entre unités de plusieurs mots

Traduction statistique

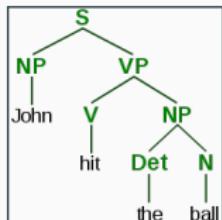
Décodage



- Construction de la phrase, de gauche à droite
- Réduction par élagage d'hypothèses faibles "beam search"

Différentes approches

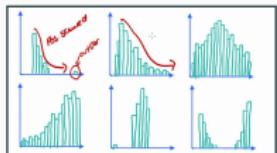
- Traduction automatique symbolique
 - basée sur la linguistique et la logique
- Traduction automatique statistique
 - basée sur le théorème de Bayes
- **Traduction automatique neuronale**
 - basée sur des réseaux de neurones



Traduction neuronale

Histoire récente

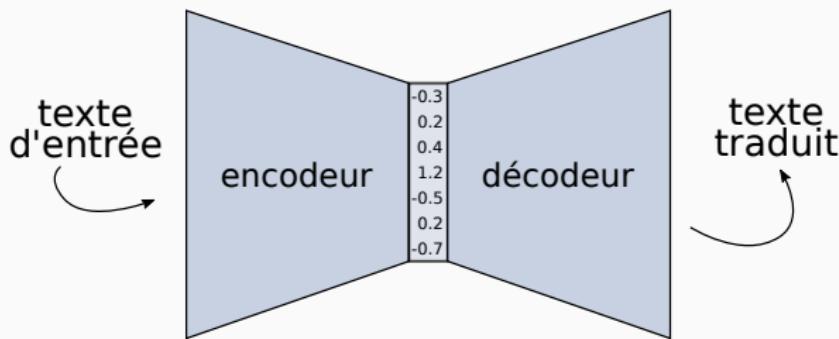
- La Traduction statistique représentait l'état de l'art jusqu'en 2015
- 2014: premier papier scientifique sur la traduction avec réseaux de neurones
- 2016: l'approche neuronale est largement adoptée comme méthode privilégiée pour la traduction automatique
- 2017: arrivée des architectures transformers



Traduction neuronale

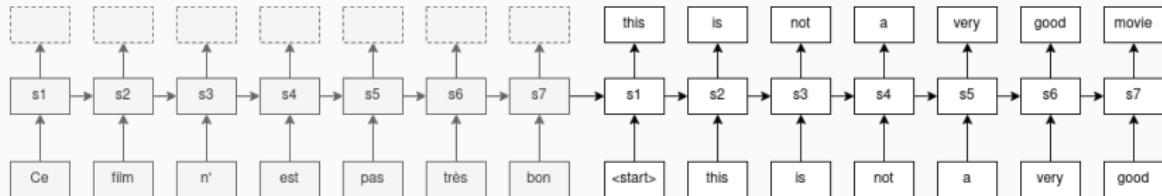
Quoi?

- L'approche neuronale consiste à modéliser le processus de traduction entier comme une grande fonction mathématique : un réseau de neurones artificiel



- Questions:
 - Comment représenter une séquence de mots comme objet mathématique ?
 - Comment représenter des mots sous forme de chiffres ?

Traduction automatique neurale



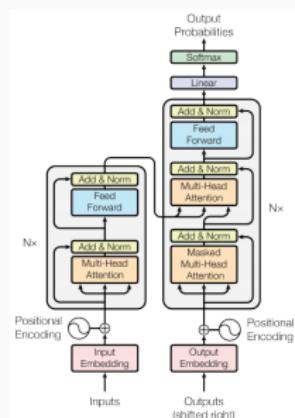
L'encodeur représente la phrase de la langue source pour conditionner le décodeur en langue cible.

Problème : goulot d'étranglement informationnel

- Toute l'entrée est encodée en une seule représentation
→ Perte d'information du début de la séquence
- Solution ? ajouter de l'information spécifique venant des éléments de l'entrée
- "Attention" → réseaux d'attention → transformers
cf cours AA1/AA2

Transformer pour la traduction ?

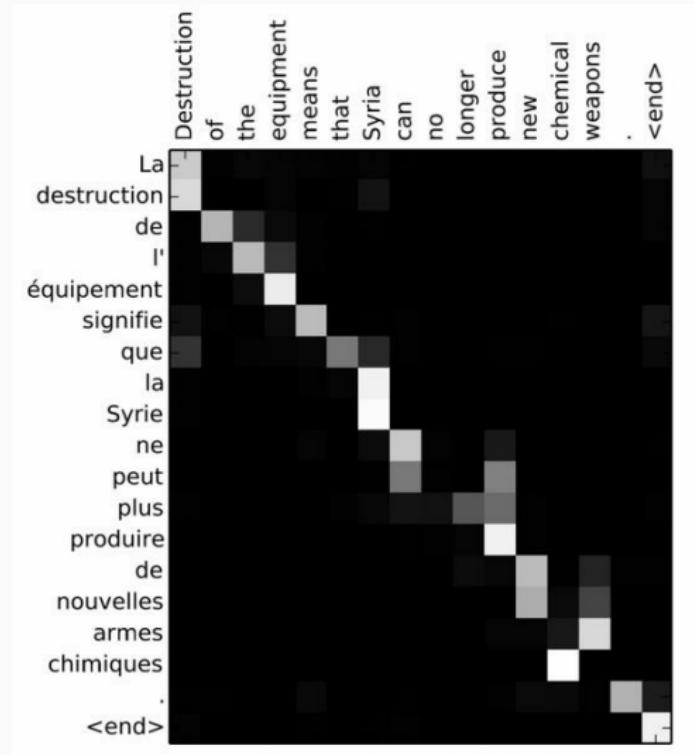
- même principe encodeur-décodeur
- ... mais plein de variantes (ex: BART, T5)
- même décodeur "pur" : modèle multilingue + instructions (GPT, BARD)



Avantages de l'attention

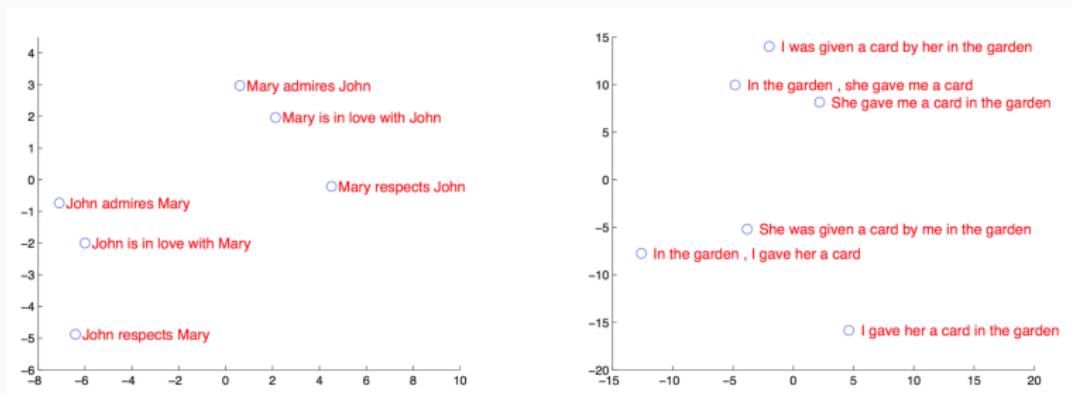
- facilite l'entraînement
- aide à l'interprétation du résultat (un peu)

Interprétation de l'alignement: traduction



Neural machine translation

Sentence representations



Entraînement

- Comment entraîner notre encodeur-décodeur ?
- On réutilise nos corpus parallèles
- Une fonction de coût quantifie l'exactitude de nos traductions par rapport à la référence
- On adapte les paramètres de notre réseau afin que notre coût soit minimal

BLEU score

- Evaluer les sorties de systèmes par des humains: couteux
- Bleu est une mesure automatique de qualité de traduction
- On compare la traduction candidate avec une traduction de référence
- On regarde
 - n-grammes (1–4) en commun avec la traduction de référence
 - Taille de la traduction
- Score de 1 = identique, 0 = aucun mot en commun
- Même des traductions par des humains n'atteindront pas un score de 1: beaucoup de façons différentes possibles de traduire une phrase

Mesures alternatives: beaucoup ! Un domaine de recherche actif.

Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie, *COMET: A Neural Framework for MT Evaluation*, ACL 2020.

Large Language Models Are State-of-the-Art Evaluators of Translation Quality (Kočmi Federmann, EAMT 2023)

Et les approches purement génératives ?

Fine-tuned vs chatGPT

	Model	Encoder	Evaluation	Team	-	-
MT (XXX→Eng)	FLoRes-200 (HRL)	ChrF++	Team et al. (2022)	63.5	-	58.64
	FLoRes-200 (LRL)	ChrF++		54.9	-	27.75
MT (Eng→XXX)	FLoRes-200 (HRL)	ChrF++	Team et al. (2022)	54.4	-	51.12
	FLoRes-200 (LRL)	ChrF++		41.9	-	21.57

A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity (Bang et al., IJCNLP-AACL 2023)

(évaluation avec ChrF++ ≈ character-level métrique + n-grams)

Neural machine translation: Problèmes non résolus

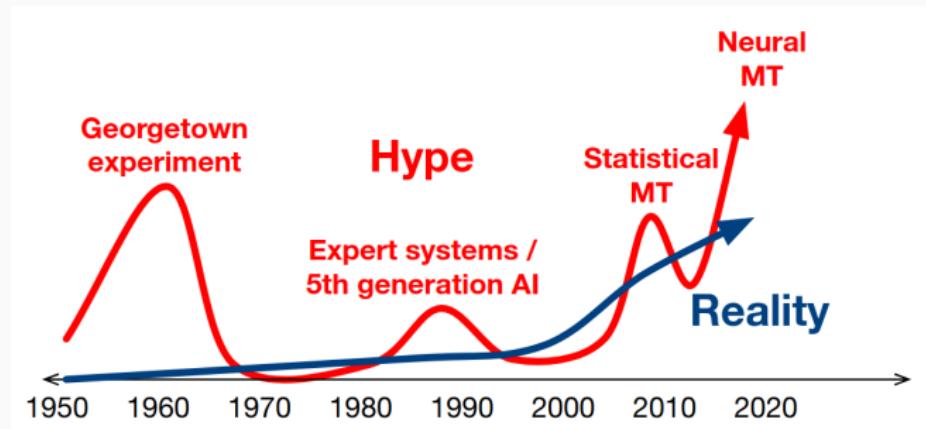
(testez avec google translate)

- mots inconnus
- langues peu dotées
- cohérence au-delà de la phrase
- résolution de pronoms; 0-pronom dans certaines langues
La thérapiste est venue s'occuper de sa patiente.
- respect du sens
- expressions figées (surtout pour des langues qui manquent de données): exemple “paper jam”
- biais:
 - The nurse came to take care of the patient.
 - The doctor came to take care of the patient.
 - The doctor came to take care of her patient.

Stabilité de la traduction ?

<https://www.translationparty.com/>

Hype ou réalité



Machine Translation: an Introduction (Koehn, 2020)
<http://mt-class.org/jhu/slides/lecture-introduction.pdf>

Dangers de la traduction automatique

