Extraction des séquences inter-langues pour la Traduction Automatique

Cyrine Nasri

LORIA MOSIC

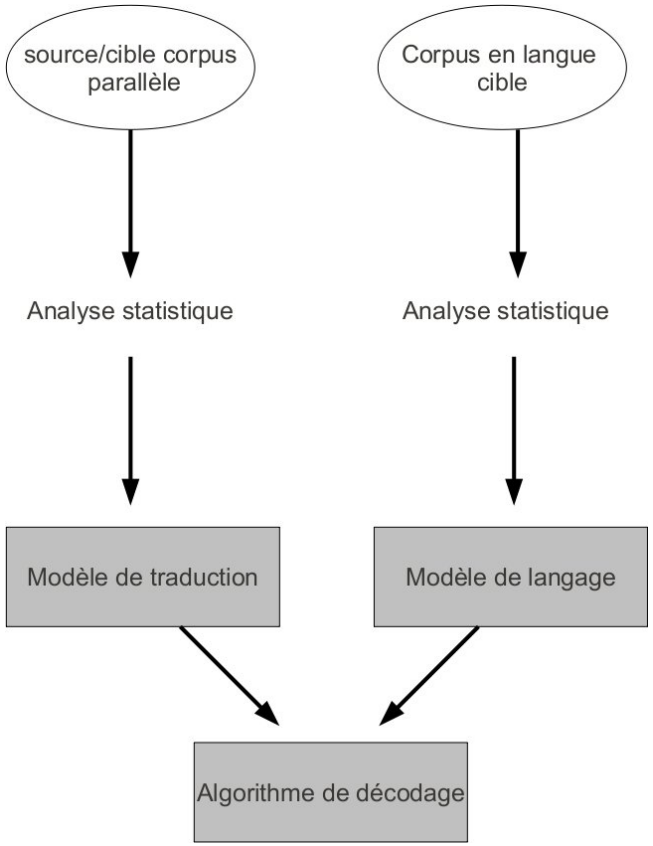
Dirigée par : Pr. Kamel Smaili, Pr. Yahya Slimani et Dr Chiraz Latiri

Plan

Introduction

- Depuis la fin de la deuxième guerre mondiale : les chercheurs se passionnent pour la TA.
- La traduction automatique : Un des premiers problèmes de l'Intelligence Artificielle.
- La traduction automatique statistique : nécessite pas l'intervention humaine.
- 50 ans de progrès :
  1. Secteur de la recherche : 1000 articles publiés (la moitié pendant les 5 dernières années).
  2. Secteur commercial : Language Weaver, Google, Microsoft.

- Composants : **Modèle de traduction**, **Modèle de langage** et le **Décodeur**



Exemple d'un corpus parallèle

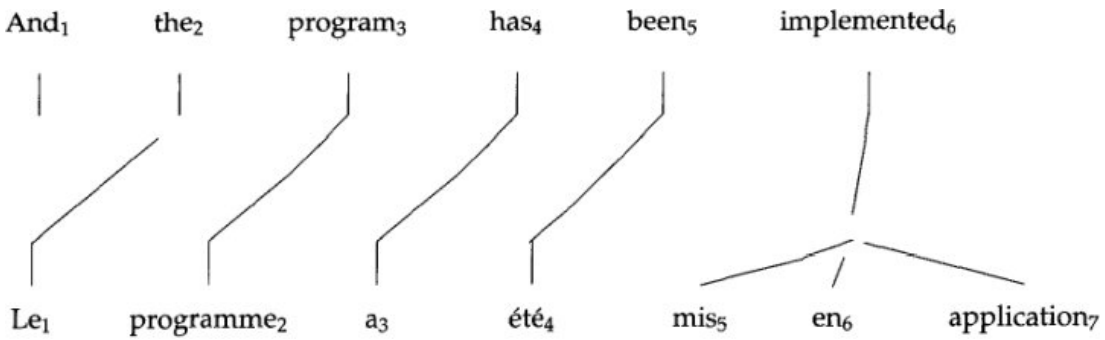
that is why the responsibility for achieving the **efficiency target** and at the same time reducing CO2 lies with the community, wich in the fact takes action when an objective can be achieved more effectively by **community measures**.	c'est pourquoi la responsabilité pour atteindre l'**objectif d'efficacité** et en même temps la réduction du CO2 incombe à la communauté, qui dans le fait prend des mesures lorsque l'objectif peut être atteint plus efficacement par des **mesures communautaires**.

- Remarque la position des mots change, le système de traduction automatique doit prendre en compte la notion de ré-ordonnement des mots.

Evaluation de la qualité la machine de traduction : Score BLEU

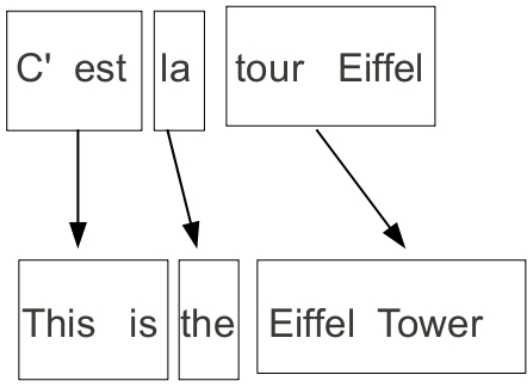
- Evaluer une traduction pose un problème en lui même.
- Introduire une métrique BLEU ...
- Mesure les similarités avec les traductions de réference.

Traduction automatique statistique à base des mots

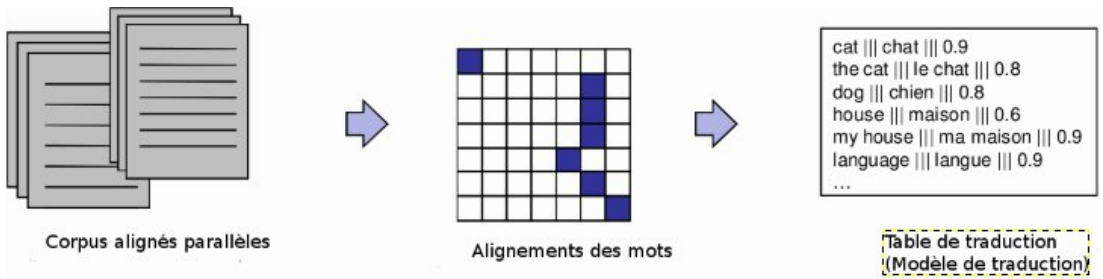


- Le processus de traduction est **décomposé en petites étapes**, *[Knight, 1997]* chacune est liée à des mots.
- Des modèles originaux pour la traduction automatique statistique : **Modèles IBM** *[Brown et al., 1993]* .

Traduction automatique statistique à base des séquences



- La phrase source et cible sont segmentés en des séquences.
- Chaque séquence source est traduite en une séquence cible.
- Processus de réordonnancement.

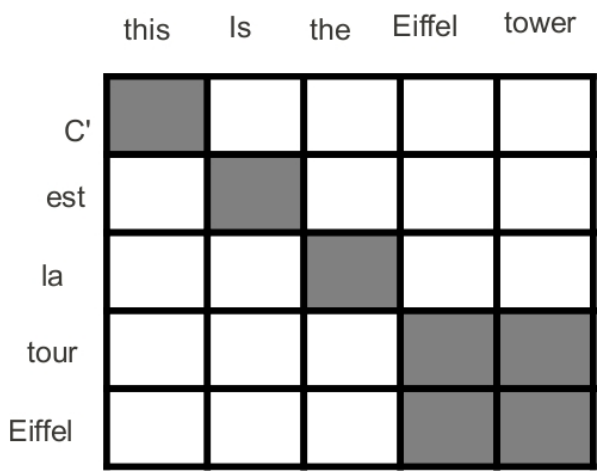


Exemple de méthodes de traduction automatique statistique à base des séquences

- Alignements (Och et al, 1999)
- Triggers inter-langues (Lavecchia et al, 2007)

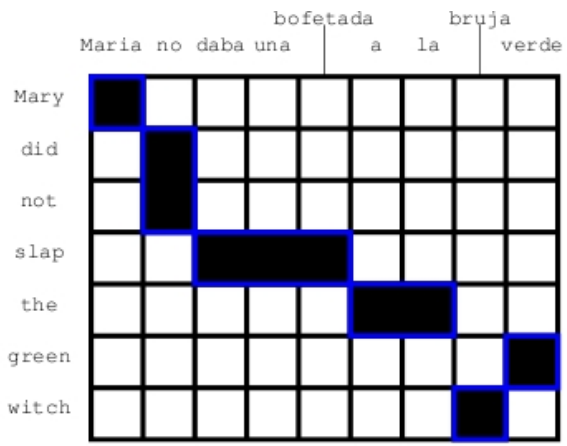
Apprentissage de la table de traduction à base de séquences (Och et al.)

- Commencer par **les alignements entre les mots**.



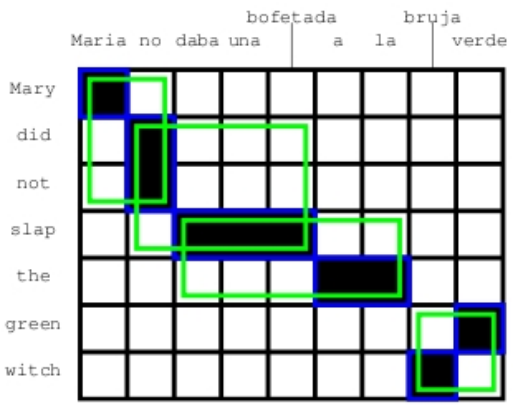
- Collecter toutes les paires de séquences qui ont des points d'alignement communs entre les mots qui icluent.

Des alignements entre les mots vers des séquences



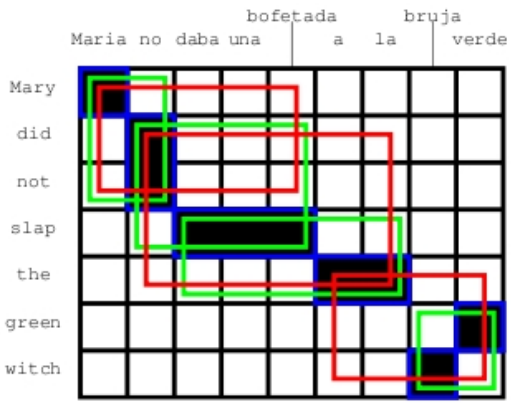
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

Des alignements entre les mots vers des séquences



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch)

Des alignements entre les mots vers des séquences



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

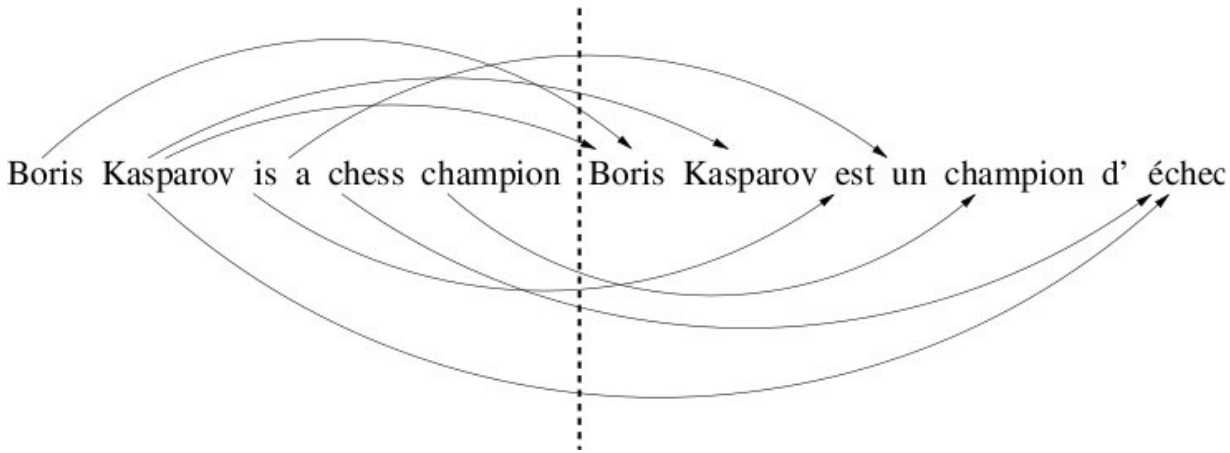
Les triggers dans la modélisation statistique du langage

- Un trigger est un ensemble composé d'un mot et ses meilleurs mots qui lui sont corrélés.
- Les triggers sont determinés par le calcul de l'Information Mutuelle (IM) entre les mots:

$$MI(a,b) = P(a,b) log \frac{P(a,b)}{P(a)P(b)} \quad (1)$$

Les triggers inter-langues

- Un trigger inter-langues est un ensemble composé d'un mot source S et ses mots cibles qui lui sont fortement corrélés $t_1 \ldots t_n$
- les triggers inter-langues sont determinés par le calcul de l'IM entre les mots d'un **corpus aligné parallèle**.



Apprentissage des traductions potentielles entre les séquences [Lavecchia et al. 2007]

1. Une séquence source peut être traduite par différentes séquences cibles de différentes tailles.
2. Heuristique: Chaque séquence source de l mots peut être traduite par une séquence de j mots cibles.
3. Calculer les triggers inter-langues entre les mots en langue source et les mots en langue cible.
4. Concaténer les triggers pour former des triggers entre des séquences en langue source et en langue cible.

Problématique

- Proposer un nouveau modèle de traduction :
  - à base des séquences.
  - Récupère automatiquement les séquences et leurs traductions directement sans passer par les alignements entre les mots.
  - Fondé sur l'information mutuelle conditionnelle.
  - Améliorer la qualité de traduction en terme de score BLEU.

$\Rightarrow$ Etendre la notion de l'information mutuelle entre deux variables vers l'information mutuelle entre trois variables : Information mutuelle conditionnelle

Principe de l'Information Mutuelle Conditionnelle

- L'information mutuelle conditionnelle entre 3 variables :

$$I(X,Y|Z) = \sum_{z\in Z}\sum_{y\in Y}\sum_{x\in X} P(x,y,z) log \frac{P(x,y,z)P(z)}{P(x,z)P(y,z)} \quad (2)$$

Algorithme : Apprentissage de notre modèle de traduction

- Calculer les triggers 2 $\longrightarrow$ 1 Francais-Anglais (X, Y : deux mots en langue source et Z: mot en langue cible).
- Reécrire le corpus Francais avec les meilleurs triggers (concaténer X, Y pour former une seule entité).
- Calculer les triggers 2 $\longrightarrow$ 1 Anglais-Francais : Pour chaque unité en garder les k meilleurs traductions :
- Reécrire le corpus Anglais avec les meilleurs triggers.

Construction de notre table de traduction

Calcul des probabilités

- $f \longrightarrow e_1 IMC = IM_1$
- $f \longrightarrow e_2 IMC = IM_2$
- $f \longrightarrow e_3 IMC = IM_3$
- $f \longrightarrow e_4 IMC = IM_4$

Table de traduction

$f \| e_1 \| \frac{IM_1}{\sum IMC(f \longrightarrow e_i)}$

Description des données

Corpus Europarl 2005		Anglais	Français
Apprentissage	Phrases	596831	596831
	Mots	15138093	16613485
	Vocabulaire	59838	76946
Développement	Phrases	1444	1444
	Mots	14077	13770
	Vocabulaire	2274	2701
Test	Phrases	500	500
	Mots	4945	5249
	Vocabulaire	1153	1352

Résultats

Table de traduction	Score Bleu
Lavecchia et al.	34.18
Och et al. 3	42.75
Notre approche	38.43

- **Cause** :L'information mutuelle ne prends en compte qu'un seul mot dans la langue cible, le reste est recupéré par concaténation.
- **Nouvelle orientation : IM qui porte sur plusieurs mots (IM Multivariables)**

Nouvelle démarche : IM Multivariables

- **Premièr étape**
  identifier les n_i meilleures séquences dans les corpus français
  identifier les n_i meilleures séquences dans les corpus anglais
- **Deuxième étape**
  Calculer les IM entre les séquence $(f_1, f_2, ..., f_n, a_1, a_2, ..., a_n)$

$$IM(f_1, f_2, ..., f_n, a_1, a_2, ..., a_m) =$$

$$P(f_1, f_2, ..., f_n, a_1, a_2, ..., a_m) log \frac{P(f_1,...,f_n,a_1,...,a_m)}{P(f_1,f_2,...,f_n),P(a_1,a_2,...,a_m)} (3)$$

- Pour chaque séquence source, on garde les k meilleures traductions

Résultats

Table de traduction	Score Bleu
Lavecchia et al.	34.18
Och et al.	42.75
Notre approche IMC	38.43
Notre approche IM Multivariables	39.42

Travaux à court et à moyen terme

- Filtrer la liste des séquences sources : ne laisser que les séquences composés des mots fortement corrélés.
- Améliorer le modèle de langage en se basant sur les sequences de mots au lieu de mots.

Merci pour votre attention!